Essays zu methodischen Herausforderungen im Large-Scale Assessment

Dissertation

zur Erlangung des akademischen Grades

Dr. phil.

im Fach Erziehungswissenschaften

eingereicht am 2. Juli 2015
an der Kultur-, Sozial- und Bildungswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von
Dipl.-Math. Alexander Robitzsch

Präsident der Humboldt-Universität zu Berlin
Prof. Dr. Jan-Hendrik Olbertz

Dekanin der Kultur-, Sozial- und Bildungswissenschaftlichen Fakultät
Prof. Dr. Julia von Blumenthal

Gutachter:
1. Prof. Dr. Olaf Köller
2. Prof. Dr. Oliver Lüdtke
3. Prof. Dr. Hans Anand Pant

Tag der Verteidigung:
27. Oktober 2015

Für Andrea

Zusammenfassung

Mit der wachsenden Verbreitung empirischer Schulleistungsleistungen im Large-Scale Assessment gehen eine Reihe methodischer Herausforderungen einher. Dabei sind Item-Response-Modelle (IRT-Modelle) zentral in der Skalierung von Schülerfähigkeiten.

Die vorliegende Arbeit untersucht, welche Konsequenzen Modellverletzungen in eindimensionalen IRT-Modellen (insbesondere im Rasch-Modell) besitzen. Insbesondere beschäftigt sich die vorliegende Arbeit mit den folgenden vier methodischen Herausforderungen von Modellverletzungen.

Erstens, implizieren Positions- und Kontexteffekte, dass gegenüber einem eindimensionalen IRT-Modell Itemschwierigkeiten nicht unabhängig von der Position im Testheft und der Zusammenstellung des Testheftes ausgeprägt sind und Schülerfähigkeiten im Verlauf eines Tests variieren können. Zweitens, verursacht die Administration von Items innerhalb von Testlets lokale Abhängigkeiten, wobei unklar ist, ob und wie diese in der Skalierung berücksichtigt werden sollen. Drittens, können Itemschwierigkeiten aufgrund verschiedener Lerngelegenheiten zwischen Schulklassen variieren. Viertens, sind insbesondere in low stakes Tests nicht bearbeitete Items vorzufinden.

Diese Arbeit liefert empirische Befunde zu den oben genannten Herausforderungen. Insbesondere wird argumentiert, dass trotz Modellverletzungen nicht zwingend von verzerrten Schätzungen von Itemschwierigkeiten, Personenfähigkeiten und Reliabilitäten ausgegangen werden muss. Außerdem wird hervorgehoben, dass man psychometrisch häufig nicht entscheiden kann und entscheiden sollte, welches IRT-Modell vorzuziehen ist. Dies trifft auch auf die Fragestellung zu, wie nicht bearbeitete Items zu bewerten sind. Ausschließlich Validitätsüberlegungen können dafür Hinweise geben.

Modellverletzungen in IRT-Modellen lassen sich konzeptuell plausibel in den Ansatz des Domain Samplings (Item Sampling; Generalisierbarkeitstheorie) einordnen. Dabei wird nicht nur auf die Existenz latenter Variablen Bezug genommen. In dieser Arbeit wird gezeigt, dass die statistische Unsicherheit in der Modellierung von Kompetenzen nicht nur von der Stichprobe der Personen, sondern auch von der Stichprobe der Items und der Wahl statistischer Modelle verursacht wird.

Abstract

Several methodological challenges emerge in large-scale student assessment studies like PISA and TIMSS. Item response models (IRT models) are essential for scaling student abilities within these studies.

This thesis investigates the consequences of several model violations in unidimensional IRT models (especially in the Rasch model). In particular, this thesis focuses on the following four methodological challenges of model violations.

First, position effects and contextual effects imply (in comparison to unidimensional IRT models) that item difficulties depend on the item position in a test booklet as well as on the composition of a test booklet. Furthermore, student abilities are allowed to vary among test positions. Second, the administration of items within testlets causes local dependencies, but it is unclear whether and how these dependencies should be taken into account for the scaling of student abilities. Third, item difficulties can vary among different school classes due to different opportunities to learn. Fourth, the amount of omitted items is in general non-negligible in low stakes tests.

This thesis provides empirical evidence for the above mentioned methodological challenges. It is argued that estimates of item difficulties, student abilities and reliabilities can be unbiased despite model violations. Furthermore, it is argued that the choice of an IRT model cannot and should not be made (solely) from a psychometric perspective. This also holds true for the problem of how to score omitted items. Only validity considerations provide reasons for choosing an adequate scoring procedure.

Model violations in IRT models can be conceptually classified within the approach of domain sampling (item sampling; generalizability theory). In this approach, the existence of latent variables need not be posed. It is argued that statistical uncertainty in modelling competencies does not only depend on the sampling of persons, but also on the sampling of items and on the choice of statistical models.

Danksagung

Die vorliegende Arbeit ist mit Unterstützung vieler Personen zustande gekommen, denen ich an dieser Stelle danken möchte.

Zunächst danke ich Olaf Köller, Oliver Lüdtke und Hans Anand Pant für die Bereitschaft, diese Arbeit zu begutachten. Durch zahlreiche Diskussionen mit Oliver Lüdtke konnte ich Unklarheiten und argumentative Schwächen in einzelnen Kapiteln der Dissertation beseitigen.

In meiner bisherigen wissenschaftlichen Laufbahn habe ich viel aus Betreuung, Kooperationen und Projekten mit Olaf Köller, Oliver Lüdtke, Marja van den Heuvel-Panhuizen und Lutz-Michael Alisch lernen und bereichernd in meine Arbeiten einbinden können.

Andrea Kruse und Matthias Trendtel danke ich für das genaue Korrekturlesen und viele hilfreiche inhaltliche Kommentare zu einzelnen Kapiteln dieser Arbeit.

Die „Rahmung“ dieser Arbeit (Kapitel 1 und 7) ist auch aufgrund der inspirierenden Arbeitsatmosphäre in den letzten Jahren am BIFIE Salzburg entstanden, die auch genügend Raum für wissenschaftliche Tätigkeiten bot.

Schließlich danke ich meiner Familie für die langjährige Unterstützung. Ich danke natürlich vor allem Andrea, die zum Erfolg in der Endphase der Arbeit maßgeblich beigetragen hat.

Inhaltsverzeichnis

1 Item-Response-Modelle: Ein kurzer Überblick **1**
- 1.1 Rolle der Personen und Items im Rasch-Modell . . . 1
- 1.2 Schätzmethoden . . . 5
- 1.3 Weitere eindimensionale IRT-Modelle . . . 9
- 1.4 Spezifische Objektivität und Skalenniveau . . . 14
- 1.5 Mehrdimensionale IRT-Modelle . . . 20
- 1.6 IRT-Modelle, log-lineare Modelle und Ising-Modell . . . 22
- 1.7 Restringierte Latent-Class-Modelle . . . 24
- 1.8 Unscharfe latente Variablen und unscharfe Item Responses . . . 27

2 Fragestellungen der Arbeit **32**
- 2.1 Kapitel 3: Ausgewählte methodische Herausforderungen bei der Kalibrierung von Leistungstests . . . 32
- 2.2 Kapitel 4: Bedeutung der Itemauswahl und der Modellwahl in Längsschnittstudien . . . 34
- 2.3 Kapitel 5: Modellierung lokaler Abhängigkeiten . . . 35
- 2.4 Kapitel 6: Item-Response-Modelle für fehlende Item Responses . . . 36
- 2.5 Kapitel 7: Abschließendes Resümee . . . 37

3 Einige methodische Herausforderungen bei der Kalibrierung von Leistungstests **39**
- 3.1 Alternativen zum Rasch-Modell . . . 39
- 3.2 Zur Interpretation der latenten Variablen im Rasch-Modell . . . 41
- 3.3 Positions- und Bookleteffekte in Large-Scale-Assessments . . . 45
  - 3.3.1 Positionseffekte auf der Itemseite . . . 46
  - 3.3.2 Positionseffekte auf der Personenseite . . . 49
  - 3.3.3 Untersuchung von Bookleteffekten . . . 51
  - 3.3.4 Schlussfolgerungen . . . 53
- 3.4 Testleteffekte: Zur Modellierung von Abhängigkeiten von Items mit einem gemeinsamen Stimulus . . . 55
  - 3.4.1 Testleteffekte für den Kompetenzbereich Lesen . . . 56
  - 3.4.2 Diskussion . . . 57
- 3.5 Multilevel DIF: Modellierung klassenspezifischer Itemschwierigkeiten . . . 59
  - 3.5.1 Multilevel DIF bei einem Rechtschreibtest der Bildungsstandards . 61
  - 3.5.2 Multilevel DIF beim Mathematiktest DEMAT . . . 61

3.5.3 Simultane Betrachtung von Testleteffekten und Multilevel DIF . . . 63
3.5.4 Diskussion . . . 65
3.6 Abschließende Bemerkungen . . . 66
3.6.1 Multilevel IRT-Modelle . . . 66
3.6.2 Zur Modellwahl . . . 66

4 Bedeutung der Itemauswahl und der Modellwahl für die längsschnittliche Erfassung von Kompetenzen 68
4.1 Einleitung . . . 68
4.2 Lesekompetenz und Lesekompetenzentwicklung . . . 69
4.3 Konzept der Generalisierbarkeit für Tests . . . 69
4.3.1 Item Sampling . . . 71
4.3.2 Zur Wahl eines IRT-Modells . . . 73
4.3.3 Modellselektion, (Multi) Model Inference, Model Averaging oder Model Sampling? . . . 74
4.3.4 Effektgrößen für Längsschnittdaten . . . 75
4.3.5 Item Parameter Drift . . . 76
4.4 Fragestellungen . . . 77
4.5 Methode . . . 77
4.5.1 Stichprobe . . . 77
4.5.2 Instrumente . . . 77
4.5.3 Behandlung fehlender Daten . . . 78
4.5.4 Statistische Analysen . . . 78
4.6 Ergebnisse . . . 81
4.6.1 Deskriptive Befunde: Effektgrößen und Stabilitäten . . . 81
4.6.2 Varianzkomponentenmodelle . . . 82
4.6.3 Item Parameter Drift . . . 82
4.6.4 Effektgrößen aus mehrdimensionalen Rasch-Modellen . . . 83
4.7 Diskussion . . . 85

5 Zur (Nicht-)Modellierung lokaler Abhängigkeiten in Messmodellen 91
5.1 Einleitung . . . 91
5.2 Psychometrische Modellierung lokaler Abhängigkeiten . . . 92
5.2.1 Drei Ansätze zur Modellierung lokaler Abhängigkeit . . . 92
5.2.2 Domain Sampling . . . 96
5.2.3 Rolle des Modellfehlers bei der Schätzung der Reliabilität . . . 99
5.2.4 Unbestimmtheit der Reliabilität und Modelläquivalenz . . . 101
5.2.5 Zwischenresümee . . . 104
5.3 HAMLET-Test . . . 104
5.3.1 Material . . . 104
5.3.2 Statistische Analysen . . . 105
5.3.3 Ergebnisse . . . 105
5.4 Bedeutung der lokalen Abhängigkeit im Kontext des Reliabilität-Validitäts-Dilemmas . . . 110
5.4.1 Ein gemeinsames Modell für Reliabilität und Validität . . . 110

5.4.2 Einschränkung der Validität durch Elimination lokaler Abhängigkeiten . . . 113
5.4.3 Zusammenfassung . . . 116
5.5 Diskussion . . . 117

6 Nichtignorierbare Konsequenzen des (partiellen) Ignorierens fehlender Item Responses im Large-Scale Assessment 122
6.1 Fehlende Item Responses in Large-Scale Assessments . . . 122
6.2 Eine Auseinandersetzung mit Kritikpunkten des „traditionellen“ Vorgehens bei fehlenden Item Responses . . . 123
6.2.1 Aleatorische und epistemische Unsicherheit . . . 125
6.2.2 Testtheoretische „Begründungen“ . . . 125
6.3 Modellbasierte Behandlung fehlender Item Responses . . . 128
6.4 Zwei alternative Item-Response-Modelle für nichtignorierbare Item Responses: Ansätze für eine Sensitivitätsanalyse . . . 130
6.4.1 Pseudo-Likelihood-Ansatz für partielles Scoring der Item Responses 130
6.4.2 Modellierung des Ausfallprozesses der Item Responses . . . 132
6.5 Ländervergleich von vier Ländern in PIRLS 2011 . . . 133
6.5.1 Daten . . . 133
6.5.2 Analysen . . . 134
6.5.3 Ergebnisse . . . 134
6.6 Diskussion . . . 135

7 Abschließendes Resümee 139
7.1 Modellierung von Positions-, Ermüdungs- und Kontexteffekten . . . 139
7.1.1 Modellabweichungen im Rasch-Modell . . . 139
7.1.2 Positionseffekte . . . 142
7.1.3 Ermüdungseffekte . . . 144
7.1.4 Bedeutung für Gruppenvergleiche . . . 145
7.1.5 Kontexteffekte und Bookleteffekte . . . 146
7.1.6 Bedeutung von Positions-, Ermüdungs- und Kontexteffekten in Längsschnittanalysen . . . 151
7.2 Bedeutung der Mehrebenenstruktur für IRT-Modelle . . . 154
7.2.1 Mehrebenenstruktur als Störquelle . . . 154
7.2.2 Mehrebenenstruktur von substanziellem Interesse . . . 157
7.2.3 Multilevel DIF: Fähigkeiten als Level-2-Konstrukt . . . 160
7.2.4 Multilevel DIF und Instruktionssensivität . . . 162
7.3 Rolle des Domain Samplings in der Item-Response-Theorie . . . 163
7.3.1 Domain Sampling . . . 163
7.3.2 Item-Response-Theorie und Generalisierbarkeitstheorie . . . 165
7.3.3 Domain Sampling Interpretation von Cronbachs Alpha . . . 169
7.3.4 Mehrdimensionale Faktormodelle . . . 173
7.3.5 Bedeutung der Invarianztestung für Gruppenvergleiche . . . 186

Literaturverzeichnis 197

Kapitel 1

Item-Response-Modelle: Ein kurzer Überblick

Dieses Kapitel gibt einen kurzen (subjektiven) Überblick über Item-Response-Modelle, die im Large-Scale Assessment Relevanz besitzen. Zunächst wird in Abschnitt 1.1 das Rasch-Modell eingeführt und verschiedene Interpretationen von Personen- und Itemparametern vorgenommen. Diese Interpretationen führen häufig zur Wahl bestimmter Schätzmethoden, die in Abschnitt 1.2 diskutiert werden. Verschiedene eindimensionale Item-Response-Modelle als Alternativen zum Rasch-Modell werden in Abschnitt 1.3 besprochen. Einige primär dem Rasch-Modell zugeschriebene Mythen werden in 1.4 kritisch beleuchtet. Abschnitt 1.5 gibt einen ausschnittsweisen Überblick zu mehrdimensionalen IRT-Modellen. Wir zeigen in Abschnitt 1.6, dass Item-Response-Modelle mit latenten Variablen äquivalent durch Verteilungen repräsentiert werden können, die ausschließlich auf manifesten Item Responses beruhen. In Abschnitt 1.7 ordnen wir die bis dahin eingeführten ein- und mehrdimensionalen Item-Response-Modelle in die fundamentale Klasse der restringierten Latent-Class-Modelle ein. Abschließend (und als Ausblick) übertragen wir diesen Ansatz in Abschnitt 1.8 auf unscharfe manifeste Item Responses und unscharfe latente Variablen.

1.1 Rolle der Personen und Items im Rasch-Modell

Rasch-Modell

Wir bezeichnen mit X_{pi} einen dichotomen Item Response (Itemantwort) für Person p auf Item i. Item-Response-Modelle sind statistische Modelle für die Menge der Zufallsvariablen X_{pi} $(p = 1, \ldots, N; i = 1, \ldots, I)$ (siehe z.B. Yen & Fitzpatrick, 2006 oder Bock & Moustaki, 2007 für einen Überblick). Eines der wichtigsten IRT-Modelle ist das *Rasch-Modell* (Rasch, 1960; siehe Fischer, 2007 für einen Überblick). Die Modellgleichung des Rasch-Modells ist durch

$$P(X_{pi} = 1) = \frac{\exp(\theta_p - b_i)}{1 + \exp(\theta_p - b_i)} = \Psi(\theta_p - b_i) \tag{1.1}$$

gegeben, wobei Ψ die logistische Linkfunktion bezeichnet. Häufig schreibt man (1.1) auch abkürzend als

$$\text{logit}\, P(X_{pi} = 1) = \theta_p - b_i \tag{1.2}$$

Anstelle der dichotomen Item Responses kann man das Rasch-Modell auch mit einem zugrunde liegenden stetigen Item Response X^*_{pi} anschreiben, worauf in Abschnitt 7.1.1 vertiefend eingegangen wird.

Mit der Definition der Modellgleichung (1.1) ist zunächst unklar, welchen Status die statistischen Parameter der *Personenfähigkeit* θ_p und der Itemschwierigkeit b_i besitzen (vgl. San Martin & De Boeck, 2015). Die Rolle von Personen und Item kann dabei als *fixed* oder *random* betrachtet werden (vgl. die Terminologie in der Multilevel-Literatur; Snijders & Bosker, 2012). Wir gehen im Folgenden auf die Konsequenzen der verschiedenen Kombinationen von fixed und random für Personen und Items ein. Die Unterscheidung folgt dabei üblichen Darstellungen in der Literatur (Lord & Novick, 1968; Holland, 1990a; De Boeck, 2008; San Martin & De Boeck, 2015).

Fixed Persons Fixed Items (FPFI)

In der Fixed Persons Fixed Items (FPFI) Perspektive ist jeder Person p eine feste Fähigkeit θ_p und jedem Item eine feste Itemschwierigkeit b_i zugeordnet. Die Wahrscheinlichkeit $\alpha_{pi} := P(X_{pi} = 1)$ im Rasch-Modell lässt sich wie folgt interpretieren: Person p löst Item i mit einer Wahrscheinlichkeit α_{pi}. In diesem Sinn werden also Item Responses X_{pi} für jede Person p und jedes Item i als probabilistisch betrachtet. Für N Personen und I Items enthält das Rasch-Modell damit insgesamt $N \cdot I$ Parameter, wovon $N \cdot I - 1$ Parameter identifizierbar (d.h. schätzbar) sind.

Die FPFI-Annahme folgt der *stochastic subject* Annahme (Holland, 1990a; vgl. auch die intraindividuelle propensity distribution in Lord & Novick, 1968), nach der Stochastizität in Item Responses durch die „Indeterminiertheit" der Antwort von Person p auf Item i zustandekommt. Man kann allerdings auch argumentieren, dass die beobachtete Itemantwort x_{pi} bei der Schätzung im Rasch-Modell durch α_{pi} vorhergesagt gesagt wird. Damit könnte die Itemantwort x_{pi} auch deterministisch sein, sie wird allerdings stochastisch modelliert.

Unter der FPFI-Annahme betrachtet man X_{pi} demzufolge als $N \cdot I$ untereinander unabhängige Zufallsvariablen, die jeweils Bernoulli-verteilt mit der Wahrscheinlichkeit $\alpha_{pi} = \Psi(\theta_p - b_i)$ sind (siehe San Martin & De Boeck, 2015).

Random Persons Fixed Items (RPFI)

In der Random Persons Fixed Items (RPFI) Perspektive werden Personen als zufällig betrachtet. Variabilität in Itemantworten kann demzufolge durch ein Sampling von Personen aus einer größeren Population geschehen. Die Wahrscheinlichkeit $\alpha_{pi} = \Psi(\theta_p - b_i)$ wird dann wie folgt interpretiert: Personen mit Fähigkeit θ_p lösen das Item i der Itemschwierigkeit von b_i mit einer Wahrscheinlichkeit von α_{pi}. Es wird also nur eine Aussage „im Mittel" getroffen. Die Menge der Personen mit Fähigkeit θ_p ist demzufolge nicht unterscheidbar. Es wird keine Aussage darüber getroffen, wie groß die Wahrscheinlichkeit einer korrekten Itemantwort für eine konkrete Person p_0 ist.

Für die Personenfähigkeiten ist demzufolge im RPRI-Ansatz eine Verteilungsannahme $\theta \sim F$ zu treffen. In Holland (1990a) wird dies als *random sampling* Annahme verstanden. Das Rasch-Modell (1.1) wird dann häufig in der Schreibweise einer Zufallsvariablen X_i

für das Item i notiert als

$$P(X_i = 1|\theta) = \Psi(\theta - b_i) \tag{1.3}$$

Die marginale Verteilung für das Item i lässt sich mit der Annahme bedingt unabhängiger Items (also der *lokalen stochastischen Unabhängigkeit*) schreiben als (San Martin & De Boeck, 2015)

$$P(X_i = 1) = \int P(X_i = 1|\theta)\mathrm{d}F(\theta) = \int \Psi(\theta - b_i)\,\mathrm{d}F(\theta) \tag{1.4}$$

In dieser Betrachtung wird deutlich, dass die Zufallsvariablen X_{pi} unabhängig und identisch nach der Zufallsvariablen X_i in (1.4) verteilt sind. Einzelne Personen p werden demzufolge nicht über deren Fähigkeiten θ_p „identifiziert". Stattdessen findet nur eine Anpassung einer gesamten Verteilung statt.

Die Behandlung gemäß Random Persons und Fixed Items lässt die Möglichkeit zu, dass die Item Responses x_{pi} deterministisch betrachtet werden (Molenaar, 1995). Wenn eine Person p ein Item i beantwortet, so existiert keine Zufälligkeit: die Person kann das Item entweder lösen oder nicht. Die mit dem Rasch-Modell abgeleitete Verteilung der Fähigkeiten θ kann man dann formal als „Zusammenfassung" bzw. Repräsentation der hochdimensionalen Kontingenztafel aller Item Responses $\boldsymbol{X} = (X_{pi})$ ansehen (von Davier, 2010). Itemschwierigkeiten b_i inferieren demzufolge auf einen das Item i charakterisierenden Parameter, der Eigenschaften in der Population aller Personen beschreibt.

Fixed Persons Random Items (FPRI)

In der Fixed Persons Random Items (FPRI) Perspektive werden Personen als fest und Items als zufällig betrachtet. Es wird daher jede Person p mit einer Fähigkeit θ_p charakterisiert. Die Wahrscheinlichkeit α_{pi} im Rasch-Modell gibt an, wie wahrscheinlich eine Person p Items mit Itemschwierigkeit b_i löst. Dadurch ist nicht wie im Fixed Person Fixed Items Ansatz gesagt, dass für ein konkretes Item i_0 die Wahrscheinlichkeit ebenso durch α_{pi} gegeben ist. Es werden also nur mittlere Aussagen für die Population aller Items getroffen.

Statistisch wird eine Verteilung G der Itemschwierigkeiten angenommen. Die konkrete Menge der I Items im Test repräsentiert dann eine größere Itempopulation. Eine Personenfähigkeit θ_p ist demzufolge an die Definition einer Itempopulation gekoppelt. Während Random Persons Fixed Items die einzelnen Items als Zufallsvariablen $X_i = (X_{pi})_p$ betrachtet, setzt man unter Fixed Persons Random Items die Zufallsvariablen $X_p = (X_{pi})_i$ ein. Beim Vergleich von RPFI und FPRI fällt also auf, dass in diesen beiden Ansätzen die Rollen von Personen und Items vertauscht werden. Für eine Modellschätzung kann man also schließen, dass RPFI die Matrix $\boldsymbol{X}$ mit Personen als Zeilen und Items als Spalten behandelt, während man unter FIRP die Matrix $\boldsymbol{X}^T$ betrachtet, in der Zeilen und Spalten vertauscht sind, wobei nun Items die Zeilen und Personen die Spalten darstellen.

Random Persons Random Items (RPRI)

In der Random Persons Random Items (RPRI) Perspektive werden Personen als zufällig und Items als zufällig betrachtet. Personen und Items sind dann jeweils Stichproben,

die auf größere Populationen inferieren (vgl. auch Cronbach & Shavelson, 2004). Die Wahrscheinlichkeit $\alpha_{pi} = \Psi(\theta_p - b_i)$ im Rasch-Modell drückt unter der RPRI-Annahme aus, wie groß die Wahrscheinlichkeit einer korrekten Lösung für Personen mit Fähigkeit θ_p und Items mit Schwierigkeiten b_i ist. Damit werden keine Aussagen für einzelne Personen oder einzelne Items getroffen. Sowohl Personen als auch Items werden daher als nicht unterscheidbar aufgefasst.

Die Personenfähigkeiten θ können damit einer Verteilung F und die Itemschwierigkeiten b einer Verteilung G folgen. Im Rasch-Modell unter der RPRI-Perspektive sind daher die Zufallsvariablen X_{pi} unabhängig und identisch nach einer Zufallsvariablen X verteilt, für die die Beziehung

$$P(X = 1) = \int \int \Psi(\theta - b) \, \mathrm{d}F(\theta) \, \mathrm{d}G(b) \tag{1.5}$$

gilt. Der RPRI ist in der Sprechweise der Mehrebenenmodelle ein kreuzklassifiziertes Zweiebenenmodelle mit zufälligen Personen- und Itemeffekten (Van den Noortgate, De Boeck & Meulders, 2003). Die RPRI-Perspektive wird auch häufig in der Generalisierbarkeitstheorie eingenommen (Brennan, 2001a).

Vergleich der vier Perspektiven

Historisch dominiert die Fixed Persons Fixed Items Perspektive (Rasch, 1960). In Large-Scale Assessments und allgemeinen wissenschaftlichen Fragestellungen beziehen sich Aussagen meistens auf eine Population von Personen, aus der eine konkrete Stichprobe von Personen für eine jeweilige Studie gezogen wurde. Daher scheint die Random Persons Fixed Items Perspektive häufig relevanter (siehe Wainer, 2010b; McDonald, 2011; Robitzsch, Freunberger, Itzlinger-Bruneforth, Breit & Schreiner, 2015). In diagnostischen Fragestellungen hingegen wird häufig auf eine Person für eine (hypothetischen) Itempopulation generalisiert, weshalb in diesen Kontexten Fixed Persons Random Items relevant scheint (De Boeck, 2008). Ich argumentiere in dieser Arbeit, dass gerade im Educational Assessment die Random Persons Random Items Perspektive bedeutsam ist und sich die Generalisierung häufig auf eine unendlich große Itempopulation bezieht.

Die vier Perspektiven FPFI, RPFI, FPRI und RPRI führen allerdings nicht nur zu verschiedenen Interpretationen der Wahrscheinlichkeiten im Rasch-Modell, sondern führen zu verschiedenen testbaren Konsequenzen. Die Fixed Persons Fixed Items (FPFI) Perspektive bedeutet, dass das Rasch-Modell für jede Person und jedes Item gelten muss. Demzufolge muss das Rasch-Modell auch für jede Subgruppe von Personen (z.B. beiden Geschlechtern) gelten. Damit muss Invarianz gegeben sein, es ist also kein differenzielles Itemfunktionieren zugelassen. Dies steht im Kontrast zur Random Persons Fixed Items (RPFI) Perspektive. Die Spezifikation des Rasch-Modells unter dieser Perspektive velangt die Definition einer Population von Personen, auf die sich die statistische Modellierung beziehen soll. Für die gesamte Population der Personen muss dann das Rasch-Modell gelten, nicht jedoch für einzelne Personen und damit auch nicht für Subgruppen von Personen, weil nur eine Verteilung von Fähigkeit in der Personenpopulation modelliert wird. Differenzielles Itemfunktionieren stellt daher unter RPFI keine Modellverletzung dar und die Prüfung der Invarianz ist keine testbare Konsequenz des Rasch-Modells. Nichtsdestotrotz

kann die Untersuchung von differenziellem Itemfunktionieren interessant sein und eine zusätzliche Forderung an das Messmodell darstellen, die dann aber über die Annahmen des Rasch-Modells hinausgeht.

In den von Steyer (1989) definierten *stochastischen Messmodellen* ist das Zufallsexperiment dadurch beschrieben, dass in einem ersten Schritt eine Person zufällig aus der Population der Personen gezogen wird und in einem zweiten Schritt für jedes der I fest gewählten Items ein probabilistischer Item Response entsteht (siehe auch Steyer & Eid, 2001). Damit soll mit einem Messmodell (dem IRT-Modell) einerseits die Verteilung der Personen beschrieben werden (random sampling), andererseits wird ein Item Response auch im Sinne des stochastic subject interpretiert. Steyer und Eid (2001) interpretieren das Rasch-Modell durch die Annahme der stochastic subject Perspektive jedoch im Sinne von Fixed Persons, so dass das Rasch-Modell für jede Person gelten muss und Invarianz gefordert wird. Allerdings kann auch der Fall eintreten, dass random sampling und stochastic subject Perspektive zugleich gelten (siehe Holland, 1990a) und daher im Rasch-Modell eine Mischung beider Prozesse angepasst wird. Dies entspricht in Faktormodellen dem Phänomenen, dass ungeklärt ist, dass im Itemresiduum sowohl wahre spezifische Varianz (Varianz aufgrund random sampling Annahme) als auch „Messfehler“ (Varianz aufgrund stochastic subject Annahme) konfundiert sind. Ohne zusätzliche Annahmen sind diese beiden Varianzquellen nicht trennbar.

1.2 Schätzmethoden

Mit den Perspektiven Fixed Persons und Random Persons sind verschiedene Schätzmethoden in der praktischen Anwendung verbunden. Unter der Maximum Likelihood basierten Methoden sind die Methoden (Pairwise) Marginal Maximum Likelihood (MML), Joint Maximum Likelihood (JML), (Pairwise) Conditional Maximum Likelihood (CML) zu unterscheiden. Wir diskutieren im Folgenden diese Schätzmethoden für das Rasch-Modell (siehe wiederum Fischer, 2007 für einen Überblick). Dabei wird auf die Schätzung von Itemparametern, Verteilungsparametern und Personenparametern eingegangen.

Marginal Maximum Likelihood (MML)

Wir gehen davon aus, dass Item Responses x_{pi} für alle Personen p und Items i vorliegen. Wir bezeichnen mit $P_i(k, \theta; b_i)$ die Item-Response-Funktion für die dichotomen Item Responses $k = 0$ und $k = 1$. Zusätzlich wird angenommen, dass die Fähigkeiten θ einer Verteilung $F = F_{\boldsymbol{\gamma}}$ mit unbekannten Verteilungsparametern $\boldsymbol{\gamma}$ folgen, wobei häufig die Normalverteilung gewählt wird. Die log-Likelihood-Funktion lässt sich mit der Annahme der lokalen stochastischen Unabhängigkeit schreiben als

$$\log L(\boldsymbol{b}, \boldsymbol{\gamma}) = \sum_p \log \left(\int \prod_i P_i(x_{pi}, \theta; b_i) \, \mathrm{d}F_{\boldsymbol{\gamma}(\theta)} \right) \tag{1.6}$$

Zu schätzen sind dann die Itemparameter $\boldsymbol{b} = (b_i)_i$ und Verteilungsparameter $\boldsymbol{\gamma}$. Die latente Fähigkeit wird bei der *Marginal Maximum Likelihood* (MML; siehe Fischer, 2007) Schätzung ausintergriert (marginalisiert). Wenn man für F eine Normalverteilung wählt,

so schätzt man im Rasch-Modell bei gleichzeitiger Schätzung aller Itemschwierigkeiten die Varianz der Verteilung von θ. Man kann jedoch F auch semiparametrisch schätzen, wobei bei I Items (I geradzahlig) allerdings nur $I/2$ Funktionale der Verteilung F schätzbar sind (de Leeuw & Verhelst, 1986; Lindsay, Clogg & Grego, 1991; San Martin, Rolin & Castro, 2013; San Martin & Rolin, 2013). In dem sog. *located latent class model* (Lindsay et al., 1991; vgl. auch Formann, 1993 und Bartolucci, 2007) können (maximal) $I/2$ Stützstellen der θ-Verteilung und deren Wahrscheinlichkeiten geschätzt werden. Die diskreten Stützstellen der θ-Verteilung können allerdings auch vorgegeben und nur deren Wahrscheinlichkeiten bestimmt werden (von Davier, 2008). Für die Anpassung des Rasch-Modells mit der MML-Schätzung ist also keine Normalverteilungsannahme notwendig.

Die Likelihood-Funktion in (1.6) geht von der lokalen stochastischen Unabhängigkeit aus und modelliert die vollständigen Item-Response-Pattern $P(\boldsymbol{x}_p)$. McDonald (1999) argumentiert, dass für eine Anpassung von IRT-Modellen häufig bivariate Informationen ausreichen. In sog. *pseudo maximum likelihood* Ansätzen (siehe Molenberghs & Verbeke, 2005) werden nicht die vollen Item-Response-Pattern, sondern nur Teilmengen von Items betrachtet. In der Methode der *pairwise marginal maximum likelihood* betrachtet man Itempaare (i, i'), so dass sich die Optimierungsfunktion schreiben lässt gemäß

$$\log pL(\boldsymbol{b}, \boldsymbol{\gamma}) = \sum_p \sum_{i,i'} \log \left(\int P_i(x_{pi}, \theta; b_i) \cdot P_{i'}(x_{pi'}, \theta; b_{i'}) \, \mathrm{d}F_{\boldsymbol{\gamma}(\theta)} \right) \tag{1.7}$$

Wählt man anstelle der logistischen Linkfunktion im Rasch-Modell die praktisch (bis auf einen Multiplikationsfaktor) äquivalente Probit-Linkfunktion (siehe Lord & Novick, 1968, S. 399), so lässt sich das Integral in (1.7) in geschlossener Form auswerten (Renard, Molenberghs & Geys, 2004). Ein ähnlicher Ansatz wird in der Software NOHARM verfolgt (McDonald, 1997). Der Vorteil der paarweisen Methode (1.7) liegt darin, dass nur Kontingenztafeln von Itempaaren als Input der Modellanpassung notwendig sind. Außerdem kann man mit der paarweisen Methode leicht Abweichungen von der lokalen stochastischen Unabhängigkeit umsetzen, ohne diese explizit modellieren zu müssen. Wenn für das Itempaar (i, i') ein von Null verschiedene Residualkorrelation erwartet wird, so kann der Term dieses Itempaares aus der Optimierung der Pseudo-Likelihood-Funktion entfernt werden und es können trotz lokaler Abhängigkeiten unverzerrte Itemparameter und Traitvarianzen geschätzt werden. Gerade in mehrdimensionalen IRT-Modellen sind paarweise Methoden computational weniger aufwändig.

Die MML-Schätzung korrespondiert mit der Annahme des random sampling von Personen und festen Items, also der RPFI Perspektive (San Martin & De Boeck, 2015). Selbst wenn die Fixed Persons Fixed Items Perspektive eingenommen werden soll, könnte man jedoch bei einer hinreichend flexiblen Spezifikation von F mit der MML-Methode zu erwartungstreuen Schätzungen der Itemparameter gelangen und in einem zweiten Schritt Personenparameter schätzen (z.B. mit WLEs; siehe Warm, 1989).

Joint Maximum Likelihood (JML)

Die *Joint Maximum Likelihood* (JML; siehe Fischer, 2007) Schätzung schätzt für jedes Item i eine Itemschwierigkeit b_i und jede Person p eine Personenfähigkeit θ_p simultan. Die

Log-Likelihood-Funktion lässt sich dann notieren als

$$\log L(\boldsymbol{b}, \theta_1, \ldots, \theta_N) = \sum_p \sum_i P_i(x_{pi}, \theta_p; b_i) \tag{1.8}$$

Zur Identifikation der Parameter muss der Mittelwert der Personenfähigkeiten auf Null gesetzt werden. Für Personen mit ausschließlich falschen oder ausschließlich richtigen Items (Personen mit *Extremscores*) existieren keine endliche Schätzung der Personenfähigkeit, weshalb sie aus der JML-Schätzung ausgeschlossen werden müssen. Als Konsequenz resultieren verzerrte Itemparameterschätzungen bei einer festen Anzahl von Items (siehe wiederum Fischer, 2007). Es wurde eine einfache Korrekturformel vorgeschlagen, die den Bias der Itemschwierigkeiten reduziert, indem die geschätzten Itemschwierigkeiten aus (1.8) mit $(I-1)/I$ multipliziert werden (Wright & Douglas, 1977). Diese Korrekturformel eliminiert aber nicht vollständig den Bias und es bleibt unklar, wie diese Formel für Multi-Matrix-Designs oder komplexe Multi-Facetten-Designs (Linacre, 1989) angewendet werden sollte. Aus diesen Gründen wird in der Psychometrie die JML-Schätzung nicht empfohlen und selten in Studien eingesetzt.

Holland (1990a) zeigt, dass man die JML-Schätzung als Approximation der MML-Schätzung ansehen kann. Die MML-Schätzung ist dabei eine bestimmte Restriktion der Verteilung F der Fähigkeiten θ der unrestringierten JML-Schätzung.

Die JML-Schätzung korrespondiert mit der Fixed Persons Fixed Items Perspektive. Wenn die Generalisierung auf eine größere Personenpopulation nicht plausibel ist, scheint diese Perspektive adäquat zu sein. Die JML-Schätzmethode besitzt zwar nachteilige statistische Eigenschaften, ist jedoch relativ einfach zu implementieren und äußerst schnell. Lando und Bertoli-Barsotti (2014) schlagen vor, die Zielfunktion (1.8) der JML-Schätzung so zu modifizieren, dass Personen mit Extremscores in der Schätzung nicht ausgeschlossen werden müssen. Dabei wird die Likelihood-Funktion derartig modifiziert, dass anstelle der suffizienten Statistik des Summenscores $S_p := \sum_{i=1}^{I} X_{pi}$ im Rasch-Modell ein adjustierter Score $S_p^{\varepsilon} = \varepsilon + (1-2\varepsilon)S_p$ als suffiziente Statistik mit einem geeigneten ε verwendet wird. Der Parameter ε kann so gewählt werden, dass praktisch unverzerrte Itemparameterschätzungen entstehen (Lando & Bertoli-Barsotti, 2014; Bertoli-Barsotti, Lando & Punzo, 2014).

In der Ökonometrie wird die FPFI Perspektive für Paneldaten mit dichotomen Daten mittels fixed effects Ansätzen diskutiert. Dabei werden Methoden entwickelt, die den Bias der JML-Schätzung durch analytisch ermittelte Korrekturformeln oder mit Resampling-Verfahren eliminieren (z.B. Arellano & Hahn, 2006, Arellano & Bonhomme, 2011; Fernández-Val, 2009). Dabei erweist sich in einer eigenen Simulationsstudie die Methode des Jackknifing von Items (Hahn & Newey, 2004) für eine Bias-Korrektur in den Itemschwierigkeiten (auch in Multi-Matrix-Designs) als äußerst erfolgreich.

Die beiden Ansätze von Lando und Bertoli-Barsotti (2014) sowie von Hahn und Newey (2004) sollten in der psychometrischen Literatur zu einer Relativierung der häufig geäußerten Ablehnung der JML-Methode führen, da mit diesen neueren Entwicklungen ohne Annahmen an die Verteilung der Fähigkeiten erwartungstreue Itemparameterschätzungen erhalten werden können.

Conditional Maximum Likelihood (CML)

Die *Conditional Maximum Likelihood* (CML; siehe Fischer, 2007 oder Rost, 2004) Schätzung eliminiert Personenparameter aus der Likelihood-Funktion. Daher müssen keine Verteilungsannahmen an θ getroffen werden, was historisch häufig mit der Eigenschaft der spezifischen Objektivität (Rasch, 1960) verbunden wird. Wenn man einen Item Response auf den Summenscore bedingt, so wird die Personenfähigkeit θ_p eliminiert, was die CML-Schätzung motiviert. Wir illustrieren die Idee für zwei Items i und i'. Dann ergibt sich

$$\begin{aligned}
& \frac{P(X_{pi}=1, X_{pi'}=0)}{P(X_{pi}+X_{pi'}=1)} \\
= & \frac{P(X_{pi}=1, X_{pi'}=0)}{P(X_{pi}=1, X_{pi'}=0)+P(X_{pi}=0, X_{pi'}=1)} \\
= & \frac{\frac{\exp(\theta_p-b_i)}{[1+\exp(\theta_p-b_i)]\,[1+\exp(\theta_p-b_{i'})]}}{\frac{\exp(\theta_p-b_i)}{[1+\exp(\theta_p-b_i)]\,[1+\exp(\theta_p-b_{i'})]}+\frac{\exp(\theta_p-b_{i'})}{[1+\exp(\theta_p-b_i)]\,[1+\exp(\theta_p-b_{i'})]}} \\
= & \frac{\exp(\theta_p-b_i)}{\exp(\theta_p-b_i)+\exp(\theta_p-b_{i'})} \\
= & \frac{\exp(-b_i)}{\exp(-b_i)+\exp(-b_{i'})}
\end{aligned} \tag{1.9}$$

In der bedingten Likelihood fallen demzufolge die Personenfähigkeiten θ_p heraus.

Die CML-Methode kann sowohl unter der Perspektive Fixed Persons als auch Random Persons eingesetzt werden. Im Fall von Fixed Persons gewinnt man ohne expliziten Rückgriff auf Personenfähigkeiten Itemparameter. Im Fall von Random Persons muss für die Bestimmung von Itemparametern keine Verteilungsspezifikation für die Fähigkeiten erfolgen. Daraus darf aber nicht die „Stichprobenunabhängigkeit“ (bezüglich der Personenstichprobe) des Rasch-Modells gefolgert werden, wie van der Linden (1994) treffend kritisiert. Die Itemparameter der CML-Schätzung sind durchaus stichprobenabhängig und besitzen nur die Eigenschaft der Konsistenz, die man auch mit MML-Schätzungen erhält.

Personenparameter kann man unter Fixierung der Itemparameter gewinnen. Allerdings führt die Verwendung individueller ML- oder WLE-Schätzungen (Warm, 1989) der Personenparameter für die Schätzung der Verteilung von θ im Allgemeinen zu verzerrten Schätzungen, falls die Messfehler der Personenparameter unberücksichtigt bleiben. Unter Annahme normalverteilter Messfehler kann man die Verteilung mit Hilfe charakteristischer Funktionen empirisch identifizieren (z. B. Delaigle & Meister, 2008). Die Varianz des Messfehlers muss dabei nicht einmal bekannt sein (Meister, 2006).

Wählt man hinreichend viele Stützstellen der θ-Verteilung in located latent class Modellen (siehe Abschnitt zu MML; Lindsay et al., 1991), so sind die MML-Schätzung und die CML-Schätzung äquivalent (siehe auch Formann, 2007 und Holland, 1990a). Daher sollte man die Bedeutung der CML-Schätzung nicht überbewerten. Außerdem ist CML

auf die Familie der Rasch-Modelle beschränkt, für 2PL-Modelle existiert keine CML-Schätzung.

Wie für die Marginal Maximum Likelihood Schätzung kann man die Idee der Conditional Maximum Likelihood Schätzung wiederum auf Paare von Items übertragen. Die bedingte Likelihood besteht dann aus Termen wie in (1.9). Pairwise Conditional Maximum Likelihood (PCML) Verfahren konvergieren relativ schnell und sind einfach zu implementieren (Zwinderman, 1995; siehe auch Fischer, 2007). Es existieren außerdem sehr einfache Varianten der PCML-Schätzung, die nichtiterativ sind, d.h. keinen Algorithmus benötigen (Choppin, 1982; Garner, 2002).

Wie in der Pairwise Marginal Maximum Likelihood Schätzung kann man bei der PCML-Schätzung bestimmte Itempaare aus der Schätzung ausschließen, wenn diese Paare mit Modellverletzungen wie lokalen stochastischen Abhängigkeiten korrespondieren.

Zusammenfassend ist die Marginal Maximum Likelihood Schätzmethode am flexibelsten und findet daher die größte Verbreitung unter den vorgestellten Schätzmethoden in der Psychometrie. In jüngerer Zeit wird jedoch die Schätzung vieler Item-Response-Modelle mit Bayesianischen Schätzverfahren vorgeschlagen (Fox, 2010). Gerade für hochparametrisierte, hierarchische oder mehrdimensionale IRT-Modelle besitzen Bayesianische Methoden Vorteile gegenüber Maximum Likelihood Ansätzen.

1.3 Weitere eindimensionale IRT-Modelle

In diesem Abschnitt verallgemeinern wir das Rasch-Modelle durch das Einfügen weiterer Itemparameter (4PL), die Schätzung alternativer Linkfunktionen und das Einfügen weiterer Personenparameter. Außerdem stellen wir das Binomialmodell als ein Testmodell dar, das keine Annahmen an die Item-Response-Funktionen stellt, sondern ein reines Sampling-Modell mit austauschbaren Items ist.

4PL-Modell

Ausgehend vom Rasch-Modell diskutieren wir im Folgenden weitere eindimensionale IRT-Modelle. Die Verallgemeinerung des Rasch-Modells besteht dann darin, neben der Itemschwierigkeit weitere Itemparameter einzuführen. Man kann das Rasch-Modell als einen Spezialfall des vierparametrischen logistischen Modells (4PL; siehe Loken & Rullison, 2010) ansehen. Die Modellgleichung des 4PL ist gegeben durch

$$P(X_{pi} = 1) = g_i + (1 - s_i - g_i) \cdot \Psi\left(a_i(\theta_p - b_i)\right) \tag{1.10}$$

Dabei ist a_i die Itemtrennschärfe, g_i der Rateparameter und s_i der Slipping-Parameter. Der Rateparameter g_i ist die Wahrscheinlichkeit, mit der Personen mit minimaler Fähigkeit das Item i (durch Raten) korrekt lösen. Der Slipping-Parameter s_i drückt den Anteil der Personen aus, der trotz maximaler Fähigkeit das Item i nicht korrekt löst. Dies könnte beispielsweise durch Flüchtigkeitsfehler oder missverständliche Itemformulieren verursacht sein. Mit $g_i = s_i = 0$ erhält man das 2PL-Modell, mit $s_i = 0$ das 3PL-Modell.

Schon für das 2PL-Modell zeigt Haberman (2005), dass man sehr schwierig von der Normalverteilung abweichende Verteilungen der Fähigkeit θ schätzen kann. Nichtsdestotrotz werden statistische Ansätze für nichtnormale Verteilungen diskutiert (siehe z.B. Woods, 2006 oder Molenaar, 2015). Selbst bei großen Stichprobenumfängen sind alle Itemparameter in (1.10) schwierig schätzbar. Demzufolge sind starke Priorverteilungen oder Gleichsetzungen von Parametern über Items hinweg für eine Stabilisierung der Schätzungen notwendig. Es erscheint aber unplausibel, eine kleine Stichprobengröße als Argument für die Verwendung des Rasch-Modells heranzuziehen (de Gruijter, 1986). Bei Einsatz eines Multiple-Choice-Tests könnte man beispielsweise ein *Difficulty+Guessing* Modell (de Gruijter, 1986; Kubinger & Draxler, 2007; Maris & Bechger, 2009) einsetzen, bei dem man im 4PL-Modell $a_i = 1$ und $s_i = 0$ setzt, so dass man nur Itemschwierigkeiten b_i und Rateparameter g_i schätzt. Eine Fixierung des Rateparameters auf $g_i = 1/M$ bei einem Multiple-Choice-Item mit M Antwortalternativen scheint ebenso denkbar (Kubinger & Draxler, 2007).

Das 4PL mit sehr großen Trennschärfen (z.B. $a_i = 10$) entspricht dem probabilistischen Guttman-Modell (Proctor, 1970). Wenn alle Items dieselbe Schwierigkeit $b_i = b$ besitzen und die Trennschärfe a_i wiederum sehr groß gewählt wird, so diskriminieren die Items nur die Bereiche $\{\theta < b\}$ und $\{\theta > b\}$, so dass man praktisch nur eine dichotome latente Variable α identifizieren kann, die gleich 1 ist, falls θ größer als b ist, ansonsten nimmt diese den Wert 0 an. Als zu schätzende Itemparameter bleiben dann nur noch der Guessing-Parameter g_i und der Slipping-Parameter s_i. Dies führt dann zum eindimensionalen DINA-Modell (Junker & Sijtsma, 2001), einem speziellen kognitiv-diagnostischen Modell (Rupp & Templin, 2008).

Aitkin und Aitkin (2011, S. 45 ff.) schlagen ein alternatives Modell zum gewöhnlich eingesetzten 3PL-Modell vor, das eine einfachere Interpretation eines möglichen Rate-Effektes bei Items erlaubt. Das vorgeschlagene *four parameter guessing model* (4PLG) ist definiert durch

$$P(X_{pi} = 1) = c_i \cdot d_i + (1 - c_i) \cdot \Psi\left(a_i(\theta_p - b_i)\right) \tag{1.11}$$

Der Parameter c_i ist dann der Anteil der Personen, die das Item i durch zufälliges Raten bearbeiten, wobei d_i der Anteil dieser Personengruppe ist, die das Item korrekt erraten. Der Anteil $1 - c_i$ löst das Item nicht durch Raten. Damit ist die Bedeutung von c_i und d_i gegenüber dem Rateparameter g_i im 3PL verschieden. Man kann im 4PLG (1.11) den Rateparameter d_i auch auf $1/M$ fixieren, wenn ein Multiple-Choice-Item M Antwortalternativen besitzt. Man stellt jedoch leicht fest, dass das 4PLG (1.11) und das 4PL (1.10) äquivalent sind. Die Parameter a_i und b_i sind in beiden Modellen identisch, die beiden anderen Parameter lassen sich wie folgt ineinander überführen:

$$c_i = g_i + s_i \ , \ d_i = \frac{g_i}{g_i + s_i} \quad \text{bzw.} \quad g_i = c_i d_i \ , \ s_i = c_i(1 - d_i) \tag{1.12}$$

Da die Itemparameter des 4PL (und damit auch des 4PLG) relativ schwierig schätzbar sind, könnte man in Anwendungen das 3PL gegen das 4PLG mit fixierten d_i-Parametern vergleichen. Wir merken an, dass das 4PLG dem 3PL entspricht, falls $d_i = 1$ gewählt wird.

Das Rateverhalten von Personen bei Multiple-Choice-Items kann jedoch auch von der Personenfähigkeit abhängen. In sog. *ability-based guessing models* werden derartige Ab-

hängigkeiten postuliert (Cao & Stokes, 2008; San Martin, Del Pino & De Boeck, 2006). Alternativ können bestimmte Distraktoren bei Multiple-Choice-Items für verschieden fähige Personengruppen verschieden attraktiv sein, was man beispielsweise in *nested logit models* (Bolt, Wollack & Suh, 2012; Suh & Bolt, 2010) spezifizieren kann.

Schätzung der Linkfunktion

In der Definition des Rasch-Modells erscheint die Wahl der logistischen Linkfunktion Ψ zunächst willkürlich. Man kann daher ein IRT-Modell mit einer zu schätzenden (monotonen und stetigen) Linkfunktion g (ein sog. *Rasch type model*; McDonald, 1999) betrachten

$$P(X_{pi}=1) = g(\theta_p - b_i) \tag{1.13}$$

Peress (2012) zeigt, dass man für unendlich viele Items in der Fixed Persons Fixed Items Perspektive die Linkfunktion g nichtparametrisch identifizieren kann. Scheiblechner (1999) schlägt eine Schätzmethode für sein sog. *ADISOP-Modell* unter der Annahme der Monotonie von g vor. Im R-Paket sirt (Robitzsch, 2015) ist die Schätzung der Linkfunktion g für die Klasse der generalisierten logistischen Linkfunktion nach Stukel (1988) implementiert. Diese Linkfunktion hängt von zwei Parametern α_1 und α_2 ab, die auch asymmetrisches Verhalten zulassen. Eigene Simulationsstudien zeigen, dass die α-Parameter der Linkfunktion unverzerrt geschätzt werden können[1]. Nur die logistische Linkfunktion liefert die Eigenschaft der Suffizienz für die Personen- und Itemparameterschätzung in der Klasse der Rasch type Modelle (1.13) (siehe Fischer, 2007). Die praktische Bedeutung der Eigenschaft der Suffizienz sollte allerdings nicht überbewertet werden, da diese die Gleichgewichtung der Items nach sich zieht (siehe Goldstein, 1980). Die Definition und Interpretation der Fähigkeit θ im IRT-Modell ist damit an die willkürliche Wahl der Linkfunktion g gekoppelt und sollte daher nicht ausschließlich psychometrisch (auf Basis eines Modellfits) begründet werden (Goldstein, 1980; siehe auch Robitzsch, Dörfler, Pfost & Artelt, 2011). Man kann auch argumentieren, dass die latente Additivität im IRT-Modell $g(\theta_p - b_i)$ eine wünschenswerte Eigenschaft ist und die Linkfunktion g so geschätzt werden soll, dass Interaktionseffekte zwischen Personen und Items im IRT-Modell so gering wie möglich ausgeprägt sind.

Binomialmodell

Das 4PL-Modell stellt eine Verallgemeinerung des Rasch-Modells dar. Nehmen wir im Rasch-Modell jedoch an, dass alle Itemschwierigkeiten b_i gleich b ausfallen, so ergibt sich das *0PL-Modell* (siehe auch Haberman, 2007)

$$P(X_{pi}=1) = \Psi(\theta_p - b) \tag{1.14}$$

Man wird einwenden, dass dieses Modell „unplausibel“ ist, da in einem Test niemals alle Items dieselbe Schwierigkeit besitzen werden. Notieren wir nun allerdings dennoch die

[1] Eine Implementierung dieses Modells ist in der Funktion `rasch.mml2` im R-Paket sirt (Robitzsch, 2015) vorgenommen.

Log-Likelihood-Funktion im JML-Ansatz (Fixed Persons Fixed Items), so folgt

$$\log L(b, \theta_1, \ldots, \theta_N) = \sum_p \sum_i P_i(x_{pi}, \theta_p; b) \tag{1.15}$$

Definieren wir eine transformierte Fähigkeit $\theta_p^* := \theta_p - b$, so ergibt sich aus (1.15) die Beziehung

$$\log L(\theta_1^*, \ldots, \theta_N^*) = \sum_p \sum_i P_i(x_{pi}, \theta_p^*; 0) = \sum_p \sum_i (x_{pi}\xi_p + (1 - x_{pi})(1 - \xi_p)) \tag{1.16}$$

wobei wir $\xi_p = \Psi(\theta_p^*)$ schreiben. Für jede Person p wird also unabhängig von den anderen Personen in der Stichprobe ein Fähigkeitsparameter ξ_p definiert. Eine Schätzung für ξ_p ist dann gerade durch den Anteil richtig gelöster Items beschrieben. Das Modell wird auch als *Binomialmodell* bezeichnet (Lord, 1965; Lord & Novick, 1968; van der Linden, 1979).

Für beobachtete Item Responses $\boldsymbol{x}_p = (x_{pi})_i$ der Person p lässt sich die Log-Likelihood im Binomialmodell mit Fixed Persons notieren als

$$\log L(\xi_1, \ldots, \xi_N) = \sum_p \log \left(\prod_i \xi_p^{x_{pi}} (1 - \xi_p)^{1 - x_{pi}} \right) \tag{1.17}$$

Man kann die Fixed Persons Perspektive in (1.17) auch durch eine Random Persons Perspektive ersetzen, in dem eine Verteilung $F = F(\xi)$ für ξ spezifiziert. Es gilt dann

$$\log L(F) = \sum_p \log \left(\int \left\{ \prod_i \xi^{x_{pi}} (1 - \xi)^{1 - x_{pi}} \right\} \mathrm{d}F(\xi) \right) \tag{1.18}$$

Ich argumentiere, dass das Binomialmodell auch ohne expliziten Rückgriff auf Itemparameter geeignet ist, die Fähigkeit einer Person in einem Test zu quantifizieren. Wir betrachten die Person p als fest. Es ist $\overline{x}_{p\bullet}$ der Anteil der richtig gelösten Items von Person p. Damit ist $\overline{x}_{p\bullet}$ ein Schätzer für ξ_p.

Wie kann man die Wahrscheinlichkeit ξ_p interpretieren? Erstens könnte ξ_p den Erwartungswert der richtig gelösten Items für Person p in einer unendlich großen Itempopulation beschreiben. Die für den Test ausgewählten Items repräsentieren dann die Itempopulation geeignet. Zweitens könnte man die Menge der Items als fest betrachten, so könnte man argumentieren, dass „zufällige Fluktuationen“ dazu führen (oder abstrakter Interaktionen von Personen und Items), dass man bei I anderen (aber strukturell ähnliche) Items nicht denselben Anteil richtig gelöster Items erhalten würde. Drittens könnte man bei I festen Items auch argumentieren, dass die Itemantworten von Person p deterministisch sind (vgl. van der Linden, 1979). Da in diesem Fall keine weitere Generalisierung erfolgen soll, ist ξ_p identisch mit $\overline{x}_{p\bullet}$. Unter der random sampling Perspektive ist man dann an der Verteilung der Personenfähigkeiten ξ_p interessiert, die häufig bei Binomialmodellen mit einer Beta-Verteilung beschrieben wird (Lord, 1965). In dieser Arbeit wird für praktische Anwendungen die erste und dritte Interpretation bevorzugt.

Diese Überlegungen sollen zeigen, dass es in einem psychometrischen Ansatz gute Gründe geben kann, Itemeigenschaften nicht explizit zu modellieren, sondern nur ein

(ggf. hierarchisches) Modell für Personen zu formulieren. Dies kann auch dadurch begründet sein, dass selbst ein komplexes IRT-Modell wie das 4PL-Modell nicht umfassend ist, Personenverhalten abzubilden. Es werden typischerweise niemals alle Personen modellkonform sein, so dass bei einer korrekten Spezifikation der Likelihood für die Interaktion aus Personen und Items neben der Personenfähigkeit θ_p weitere Personenvariablen bedeutsam sein können. Wir gehen im nächsten Abschnitt auf solche IRT-Modelle ein.

Modelle mit weiteren Personenparametern

Wenn man das Rasch-Modell als Basismodell der IRT-Modelle definiert, so wurden beim diskutierten 4PL-Modellen Parameter auf der Itemseite eingeführt, um eine bessere statistische Anpassung des Datensatzes mit den Item Responses zu erhalten. Geht man vom 1PL (dem Rasch-Modell) zum 2PL-Modell über, so wird eine Itemtrennschärfe a_i eingeführt, die angibt, dass manche Items besser oder schlechter bezüglich der Fähigkeit diskriminieren. Dadurch erhält man keine parallelen Item-Response-Funktionen mehr wie im Rasch-Modell (Rost, 2004) und damit ist die Schwierigkeitsreihenfolge der Items für Personen mit verschiedenen Fähigkeiten nicht mehr identisch. Es kann also sein, dass bestimmte Items leistungsschwächeren Schülern relativ leicht fallen und leistungsstärkeren Schülern jedoch im Vergleich zu anderen (etwas schwierigeren) Items relativ schwierig fallen. Diese Überlegung kann man jedoch auch auf die Personenseite übertragen. Es wird Schüler geben, deren Lösungsverhalten bei leichten Items sich kaum vom Verhalten bei schwierigen Items unterscheidet. Es könnte aber auch der Fall eintreten, dass das Lösungsverhalten für manche Schüler praktisch durch ein Guttman-Modell beschrieben werden kann, wobei alle leichten Items richtig gelöst und alle schwierigen Items falsch gelöst werden.

Man wird in praktischen Anwendungen daher immer Personen finden, die kein modellkonformes Verhalten aufweisen. Solche Personen können mit Methoden des *Person Fit* identifiziert werden (Meijer & Sijtsma, 2001; Emons, Sijtsma & Meijer, 2005). Das Rasch-Modell könnte dann um einen weiteren Personenparameter α_p ergänzt werden, der das Ausmaß der Modellkonformität abbildet (Conijn, Emons, van Assen & Sijtsma, 2011; Ferrando, 2007; Reise, 2000; Strandmark & Linn, 1987)

$$P(X_{pi} = 1) = \Psi\left(\alpha_p(\theta_p - b_i)\right) \tag{1.19}$$

Typischerweise wird man den Mittelwert der Variablen α_p auf Eins fixieren. Personen mit kleinen „Personentrennschärfen“ α_p nahe von Null diskriminieren schlechter zwischen leichten und schwierigen Items, Personen mit α_p-Werten deutlich größer als Eins diskriminieren besser als man über das Rasch-Modell vorhersagen würde. Man kann in einem Modell auch simultan Personen- und Itemtrennschärfen schätzen, so dass die Modellgleichung dann als

$$P(X_{pi} = 1) = \Psi\left(\alpha_p a_i(\theta_p - b_i)\right) \tag{1.20}$$

gegeben ist (Strandmark & Linn, 1987; Raiche, Magis, Blais & Brochu, 2013). Es ist denkbar, dass Guessing-Verhalten und Slipping-Verhalten nicht nur itemspezifisch, sondern auch personenspezifisch ausfallen. Entsprechende IRT-Modelle werden in Raiche et al. (2013) diskutiert (vgl. auch Formann & Kohlmann, 1998).

1.4 Spezifische Objektivität und Skalenniveau

Gerade aus einer historischen Perspektive ranken sich um die besondere Rolle des Rasch-Modells in der psychometrischen Literatur Mythen, die dem Rasch-Modell (unberechtigterweise) gegenüber anderen Modellen einen besonderen Status beimessen (siehe Goldstein, 1980 oder McDonald, 1999 für eine Kritik). Im Folgenden werden für die Aspekte der sog. spezifischen Objektivität und des Skalenniveaus Eigenschaftszuschreibungen des Rasch-Modells von zwei häufig zitierten deutschsprachigen Lehrbüchern (Bühner, 2006; Strobl, 2010) diskutiert. Die Analyse zeigt, dass die Aussagen in dieser Literatur geeignet scheinen, die Mythen des Rasch-Modells weiter zu verbreiten oder – positiver formuliert – zumindest nicht zu widerlegen.

Begriff der spezifische Objektivität

Ehe wir näher auf Aussagen der Lehrbuchliteratur eingehen, führen wir die Definition der spezifischen Objektivität im Sinne von Rasch (1960) ein und verallgemeinern diese nachfolgend.

Die nachfolgende Definition der spezifischen Objektivität im Sinne von Rasch findet man in Fischer (2007, S. 528ff.). Wir bezeichnen die Menge der Personen mit $\mathcal{P}$ und die Menge der Items mit $\mathcal{I}$. Jeder Person $p \in \mathcal{P}$ möge ein Personenparameter ξ_p und jedem Item $i \in \mathcal{I}$ ein Itemparameter δ_i zugeordnet sein. Die Zufallsvariable X_{pi} für den Item Response für Person p auf Item i ist durch ein Reaktionsparameter $\rho_{pi} = P(X_{pi} = 1|p, i)$ charakterisiert. Für diesen Parameter möge $\rho = G(\xi, \delta)$ gelten, wobei $G : \mathbb{R}^2 \to \mathbb{R}$ stetig und monoton wachsend im ersten Argument sowie monoton fallend im zweiten Argument ist. Zusätzlich sind $G(\xi, \cdot)$ sowie $G(\cdot, \delta)$ bijektive (d.h. eineindeutige) Funktionen. Die spezifische Objektivität im Sinne von Rasch verlangt die Existenz einer *Komparatorfunktion*[2] $K : \mathbb{R}^2 \to \mathbb{R}$, die im ersten Argument monoton wachsend und im zweiten Argument monoton fallend ist. Die *spezifische Objektivität* fordert, dass

$$K(\rho_{pi}, \rho_{qi}) = K(G(\xi_p, \delta_i), G(\xi_q, \delta_i)) =: V(\xi_p, \xi_q) \tag{1.21}$$

unabhängig von allen Items i sein soll, so dass ein „Vergleich“ mit der Funktion V nur auf Basis von Personenparametern ξ_p und ξ_q möglich ist. Diese Eigenschaft sichert, Personen unabhängig von Items vergleichen zu können. Mathematisch muss man also den Itemparameter δ_i für den Vergleich zweier Personen eliminieren können.

Mit diesen Annahmen kann gezeigt werden, dass stetige, monotone und bijektive Funktionen $\Theta, B, g, h : \mathbb{R} \to \mathbb{R}$ existieren, so dass folgende Gleichungen erfüllt sind

$$K(\rho_{pi}, \rho_{qi}) = V(\xi_p, \xi_q) = h(\Theta(\xi_p) - \Theta(\xi_q)) = h(\theta_p - \theta_q) \tag{1.22}$$

$$G(\xi_p, \delta_i) = g(\Theta(\xi_p) - B(\delta_i)) = g(\theta_p - b_i) \quad \text{mit} \quad \theta_p := \Theta(\xi_p) \text{ und } b_i = B(\delta_i) \tag{1.23}$$

Die Gleichung (1.23) führt bei Gültigkeit der spezifischen Objektivität also zum Item-Response-Modell der Form $P(X_{pi} = 1) = g(\theta_p - b_i)$ mit einer stetigen, monoton wachsenden Linkfunktion g. Damit ist spezifische Objektivität äquivalent zu latenter Additivität (siehe auch van der Linden, 1994). Erst durch die zusätzliche Forderung der Existenz einer

[2] Tutz (1989) spricht von *Vergleichsfunktion*.

suffizienten Statistik für die Personenfähigkeit oder die Existenz der CML-Schätzung erfüllt nur das Rasch-Modell mit der logistischen Linkfunktion g diese Eigenschaft (Fischer, 2007). Die Definition der spezifischen Objektivität basiert auf der Fixed Persons Fixed Items Perspektive, da für jede Person und jedes Item statistische Parameter definiert werden.

Damit ist klar, dass die Rasch type models $P(X_{pi} = 1) = \rho_{pi} = g(\theta_p - b_i)$ spezifisch objektiv sind. Definieren wir die Komparatorfunktion $K(x, y) := \mathcal{L}(x) - \mathcal{L}(y)$ unter Verwendung der logistischen Funktion $\mathcal{L}(x) := \text{logit}(x) = \Psi^{-1}(x)$, so folgt $K(\rho_{pi} - \rho_{qi}) = \theta_p - \theta_q$, also die Unabhängigkeit vom Itemparameter i.

Ein allgemeineres Konzept der spezifischen Objektivität

Irtel (1995b, S. 69ff.) verallgemeinert das Konzept der spezifischen Objektivität. In diesem Ansatz müssen Komparatorfunktionen existieren, die Personenparameter unabhängig von Itemparametern darstellen. Irtels Ansatz entspricht dem Konzept der Trennbarkeit in Tutz (1989). Im Gegensatz zum Ansatz von Rasch wird nun nicht mehr nur noch von einer Komparatorfunktion K mit zwei Argumenten, sondern mit n Argumenten ausgegangen, also $K : \mathbb{R}^n \to \mathbb{R}$. Die Funktion $K_{\mathcal{P}}(\rho_{p_1 i_1}, \rho_{p_2 i_2}, \ldots)$ heißt *spezifisch objektive Komparatorfunktion für* $\mathcal{P}$, falls der Funktionswert unabhängig von den Items $i_1, i_2, \ldots$ ist. Analog heißt die Funktion $K_{\mathcal{I}}(\rho_{p_1 i_1}, \rho_{p_2 i_2}, \ldots)$ *spezifisch objektive Komparatorfunktion für* $\mathcal{I}$, falls der Funktionswert unabhängig von den Personen $p_1, p_2, \ldots$ ist. Es ist zu betonen, dass Irtel keine Stetigkeit der Komparatorfunktion K fordert.

Irtel (1995b, S. 71ff.) zeigt, dass spezifisch objektive Messungen auf einem Ordinalskalenniveau für Modelle der nichtparametrischen IRT (Sijtsma & Molenaar, 2002) möglich sind. Nichtparametrische IRT-Modelle nach Irtel und Schmalhofer (1982) besitzen die Form $P(X_{pi} = 1) = G(\theta_p, b_i)$, wobei G doppelt monoton ist, d.h. $G(\theta_p, b_i) \geq G(\theta_q, b_i)$ genau dann, wenn $\theta_p \geq \theta_q$ sowie $G(\theta, b_i) \geq G(\theta, b_j)$ genau dann, wenn $b_i \leq b_j$. Dieses Modell wird nach Scheiblechner (1995) auch *ISOP-Modell* genannt. Definiert man die Komparatorfunktion $K_{\mathcal{P}}$ für Personen durch $K_{\mathcal{P}}(G(\theta_p, b_i), G(\theta_q, b_i)) = \mathbf{1}_{\{\theta_p \geq \theta_q\}}$ (wobei $\mathbf{1}_A$ die Indikatorfunktion für eine Menge A bezeichnet), so ist das Ergebnis unabhängig vom Item i. Das nichtparametrische IRT-Modell besitzt also Item-Response-Funktionen, die sich nicht schneiden (d.h. parallele Item-Response-Funktionen), was die Unabhängigkeit des Vergleiches zweier Personen vom Item sichert. Analog kann man zeigen, dass Items unabhängig von Personen vergleichbar sind und daher spezifische Objektivität im Sinne von Irtel gegeben ist.

Irtel (1995b, S. 72ff.) zeigt außerdem die spezifische Objektivität (im Sinne seiner Definition) für das 2PL-Modell (siehe auch Irtel, 1995a). Mit $P(X_{pi} = 1) = \Psi(a_i(\theta_p - b_i))$ definieren wir als Komparatorfunktion für Personen

$$K_{\mathcal{P}}(\rho_{p_1 i}, \rho_{p_2 i}, \rho_{p_3 i}) = \frac{\mathcal{L}(\rho_{p_1 i}) - \mathcal{L}(\rho_{p_2 i})}{\mathcal{L}(\rho_{p_3 i}) - \mathcal{L}(\rho_{p_2 i})} = \frac{\theta_{p_1} - \theta_{p_2}}{\theta_{p_3} - \theta_{p_2}} \tag{1.24}$$

wobei $\mathcal{L}(\rho_{pi}) = a_i(\theta_p - b_i)$. Im Gegensatz zum Rasch-Modell sind zum Vergleich von Personen nun allerdings drei Personen anstelle von zwei Personen notwendig. Irtel (1995b, S. 72ff.) zeigt weiter, dass auch die beiden Itemparameter im 2PL-Modell unabhängig von den Personenparameter dargestellt werden können.

Spezifische Objektivität in der Lehrbuchliteratur

Der Begriff der sog. *spezifischen Objektivität* wird sehr häufig mit dem Rasch-Modell verbunden (z.B. Bond & Fox, 2001). Dabei wird in der (angewandten Lehrbuch-) Literatur häufig gar nicht klar definiert, was unter diesem Begriff genau verstanden wird. Aus dieser Situation ist vielleicht die Verbreitung einiger Mythen des Rasch-Modells begründet.

Strobl (2010, S. 20) definiert den Begriff der spezifischen Objektivität wie folgt.

> Die so genannte spezifische Objektivität gewährleistet, dass Aussagen über die Fähigkeiten zweier Personen nicht davon abhängen, anhand welcher Aufgabe sie verglichen werden.

Hier wird also gefordert, dass es für zwei verschiedene Personen p und q egal ist, welche der Items für den „Vergleich“ herangezogen werden. Die Definition legt nahe, dass sie unter der Fixed Persons Fixed Items Perspektive erfolgt, da nicht von Gruppen von Personen gesprochen wird. Ist die Person q fähiger als die Person p, so soll die Lösungswahrscheinlichkeit $P(X_{qi} = 1)$ für alle Items i größer als $P(X_{pi} = 1)$ ausfallen (Strobl, 2010, S. 20). Dann zeigt Strobl, dass spezifische Objektivität im Rasch-Modell gegeben ist. Dazu wird gezeigt, dass der Vergleich zweier Personen p und q mit Hilfe der Lösungswahrscheinlichkeiten im Rasch-Modell unabhängig vom Item ist. Es ist $\Psi^{-1}\left(P(X_{pi} = 1)\right) = \theta_p - b_i = \operatorname{logit} P(X_{pi} = 1)$ der Logit der Lösungswahrscheinlichkeit. Dann kann man notieren (siehe Strobl, 2010, S. 21)

$$\Psi^{-1}\left(P(X_{pi} = 1)\right) - \Psi^{-1}\left(P(X_{qi} = 1)\right) = (\theta_p - b_i) - (\theta_q - b_i) = \theta_p - \theta_p \tag{1.25}$$

Bei diesem Vergleich fallen Itemparameter heraus und daraus folgt die behauptete „Itemunabhängigkeit“. Wie oben geziegt wurde, erfüllt aber jede andere monotone Linkfunktion g mit $P(X_{pi} = 1) = g(\theta_p - b_i)$ (Rasch type models) diese Eigenschaft (insbesondere auch asymmetrische Linkfunktionen; siehe Goldstein, 1980 oder McDonald, 1999). Spezifische Objektivität im Sinne der paarweisen Vergleiche von Personen bedeutet also nichts weiter als die Forderung nach einem IRT-Modell mit additiven Personen- und Itemparametern (van der Linden, 1994). Wenn also behauptet wird, dass „die spezifische Objektivität ... nur durch die parallelen ICCs im Rasch-Modell gewährleistet“ (Strobl, 2010, S. 22), so kann dies auch durch parallele ICCs in Rasch type models folgen. Die einzige Eigenschaft, die das Rasch-Modell besitzt, ist die Existenz der suffizienten Statistik des Summenscores für die Personenfähigkeit, die die Conditional Maximum Likelihood Schätzung ermöglicht (van der Linden, 1994). Häufig wird diese Eigenschaft zusätzlich für das Konzept der spezifischen Objektivität gefordert, um „stichprobenunabhängige“ Schätzungen der Itemparameter zu erhalten, weil in der CML-Schätzung Personenparameter eliminiert werden.

Strobl (2010) interpretiert das Rasch-Modell ausschließlich in der Fixed Persons Fixed Items Perspektive, so dass als Konsequenz differenzielles Itemfunktionieren eine Modellverletzung darstellt.

> Stellt sich eine Aufgabe für unterschiedliche Personengruppen als unterschiedlich schwer heraus, [... so ist diese ...] Aufgabe für den Vergleich von Personen ungeeignet und sollte aus dem Test entfernt werden (Strobl, 2010, S. 23).

Alternative Interpretationen des Rasch-Modells außer der (historischen stochastic subject) Perspektive werden nicht von Strobl präsentiert (siehe Wainer, 2010b für eine Kritik).

Bei der Einführung des 2PL-Modells merkt Strobl (2010, S. 51) an, dass keine spezifische Objektivität (im Sinne der Definition von Strobl) gegeben ist, da sich die Item-Response-Funktionen schneiden. Allerdings wird eine klare Präferenz für das Rasch-Modell und dessen Prüfung der Eigenschaften geäußert.

> Ist das Ziel, einen neuen Test zu entwickeln, dann macht es Sinn, solange nach geeigneten Aufgaben zu suchen (und ungeeignete auszusortieren), bis der Test den strengen Anforderungen des Rasch-Modells genügt, weil man dann auch von den guten Messeigenschaften des Rasch-Modells profitieren kann (Strobl, 2010, S. 51)

Es bleibt unklar, inwiefern nur das Rasch-Modell „gute Messeigenschaften“ liefern kann. Die Eigenschaft der spezifischen Objektivität (im Sinne von Strobl) hat – wie wir gleich in der Diskussion zu Bühner (2006) näher erläutern – nichts damit zu tun, dass Gruppenvergleiche nicht durchgeführt werden können, weil vielleicht im 2PL-Modell die Unabhängigkeit von den verwendeten Items nicht mehr gegeben sei. Eventuell liegen „gute Messeigenschaften“ darin begründet, dass man mit dem Rasch-Modell mindestens intervallskalierte Messungen erhalten würde (Strobl, 2010, S. 24). Wir gehen im nächsten Abschnitt kritisch auf diese Behauptung ein.

Eine leicht verschiedene Definition spezifischer Objektivität verwendet Bühner (2006, S. 315).

> [...] Spezifische Objektivität [... ist durch ...] zwei Arten invarianter Vergleiche gekennzeichnet. (1) Vergleiche zwischen Personen sind invariant über die spezifischen Items [...], und (2) Vergleiche zwischen Items sind invariant über die spezifischen Personen, an denen die Items kalibriert werden.

Bei dieser Definition fällt auf, dass im Gegensatz zur Definition von Strobl (2010) nicht von Aussagen über *zwei* Personen gesprochen wird und der Aspekt des Paarvergleiches bei Bühner (2006) demzufolge nicht betont wird.

Bühner (2006) argumentiert, dass das Rasch-Modell prüft, ob Rohwerte zu Summenscores zusammengefasst werden „dürfen“. Es ist klar, dass bei Gültigkeit des Rasch-Modells der Summenscore eine suffiziente Statistik für die Fähigkeit darstellt. Die Legitimation der Verwendung des (ungewichteten) Summenscores mit dem Rasch-Modell bedeutet aber, dass man erst durch die Anpassung eines geeigneten IRT-Modells und eine daraus gewonnene Fähigkeitsschätzung „legitime“ Fähigkeitsschätzungen erhalten würde (vgl. auch Kubinger & Draxler, 2007). Ich schließe mich dieser Sichtweise nicht an (vgl. auch Brennan, 2001b und Reise, Moore & Haviland, 2010). Nach Bühner (2006, S. 321) ist im 2PL-Modell die „Basis für spezifisch objektive Vergleiche“ verletzt. Parameterschätzungen im 2PL sind demzufolge im Unterschied zum Rasch-Modell „stichprobenabhängig“. Diese Behauptung trifft allerdings nicht nur auf das Rasch-Modell zu. Gilt ein 2PL in einer Population (von Personen), so kann man mit einer konkreten Stichprobe mit dem 2PL genauso wie dem Rasch-Modell unverzerrte Itemparameterschätzungen erhalten. Meint die „Stichprobenunabhängigkeit“ die häufig angesprochene „Itemunabhängigkeit“, so ist dies auch falsch. Wenn das 2PL für eine Menge von Items gilt und man wählt eine Teilmenge von Items, so bleiben die Itemparameterschätzungen (bei großen Personenstichproben)

wie im Rasch-Modell identisch. Man erkennt, dass Vergleiche zwischen Personengruppen immer „objektiv“ (d.h. invariant sind), wenn die Itemparameter in beiden Gruppen invariant sind und das IRT-Modell gültig ist. Dass sich die Item-Response-Funktionen im 2PL-Modell schneiden, hat nichts mit der Identifikation von Gruppenunterschieden zu tun. In diesem Sinne ist damit die Verwendung des Konzepts der spezifischen Objektivität überflüssig und sogar schädlich, da Fehlvorstellungen assoziiert werden. In Bühner (2006, S. 320) findet jedoch eine Vermischung der beiden Konzepte statt:

> Abschließend lässt sich feststellen, dass Messwerte nicht nur Intervallskalenniveau implizieren, sondern auch bestimmte Invarianzeigenschaften besitzen müssen: Unterschiede zwischen Itemparametern dürfen nicht von der verwendeten Personenstichprobe abhängen und Unterschiede zwischen Personenparametern nicht von der verwendeten Itemstichprobe (spezifische Objektivität).

Das Zitat impliziert eine Gleichheit von spezifischer Objektivität mit Invarianzeigenschaften. Wenn das korrekt wäre, wäre jedes invariante Messmodell spezifisch objektiv. Die spezifische Objektivität im Sinne des Rasch-Modells scheint aber eine andere Bedeutung zu haben (eher im Sinne von Strobl, 2010), denn ansonsten käme Bühner (2006, S. 340) nicht zum Schluss, dass das 2PL- und das 3PL-Modell „keine spezifisch objektiven Vergleiche“ im Gegensatz zum Rasch-Modell ermöglichen würden, weshalb das Rasch-Modell vorzuziehen sei. Wie oben gezeigt, ergeben sich auch für das 2PL-Modell spezifisch objektive Vergleiche, wenn man anstelle von Paaren von Items bzw. Personen auf Tripel von Items übergeht (siehe auch Steyer & Eid, 2001 für ein ähnliches Prinzip für das Modell tau-kongenerischer Messungen). „Spezifisch objektive Vergleiche“ sind nach dieser Logik immer herstellbar, wenn man ein System von Funktionen aufstellen kann, um Itemparameter unabhängig von Personenparametern mathematisch separieren zu können. Damit ist das Konzept der spezifischen Objektivität mit der Identifizierbarkeit statistischer Parameter verwandt, in der zu schätzende Modellparameter als Funktion der Wahrscheinlichkeiten ρ_{pi} dargestellt werden (San Martin, González & Tuerlinckx, 2009; San Martin et al., 2013; San Martin, González & Tuerlinckx, 2015).

Auch der CML-Schätzung wird in Bühner (2006) unberechtigterweise ein Mythos zugeschrieben, wenn behauptet wird, „dass die Itemparameterschätzung nicht mehr systematisch von der Stichprobenzusammensetzung abhängen“ (Bühner, 2006, S. 338). Die CML-Schätzung ist im gleichen Maße stichprobenabhängig wie die MML-Schätzung (siehe van der Linden, 1994). Wie in Abschnitt 1.2 gezeigt wurde, sind beide Schätzmethoden bei geeigneter Verteilungsspezifikation sogar äquivalent. Inwiefern dann nur die CML-Schätzung „Grundlage spezifisch objektiver Vergleiche“ (Bühner, 2006, S. 340) sein kann, bleibt rätselhaft.

Verhältnis von Item-Response-Theorie (IRT) und klassischer Testtheorie (KTT)

Die in diesem Abschnitt diskutierten Lehrbücher versuchen, die KTT gegenüber der IRT zu konstrastieren. Strobl (2010) merkt an:

> Die Überprüfbarkeit des Rasch-Modells [...] unterscheidet den Ansatz der [IRT] [...] von dem der klassischen psychologischen Testtheorie (vgl. z.B. [...] Steyer & Eid, 2001; Bühner, 2006 [...]) (Strobl, 2010, S. 2).

Zunächst muss definiert werden, was unter der klassischen Testtheorie (KTT) verstanden werden soll. Versteht man darunter aber die stochastischen Messmodelle im Sinne von Steyer und Eid (2001), so fallen die Modelle der tau-parallelen, tau-äquivalenten und tau-kongenerischen Messungen in die Klasse der KTT. Diese Faktormodelle sind natürlich wie das Rasch-Modell im Gegensatz zur Behauptung von Strobl (2010) überprüfbar. Damit besteht in der Terminologie nach Steyer und Eid (2001) kein formaler Unterschied zwischen KTT und IRT. Ich argumentiere allerdings, dass die KTT eher als Sampling-Modell für (austauschbare) Items als ein Messmodell für feste Items angesehen werden sollte (siehe Brennan, 2011 oder Cronbach & Shavelson, 2004).

Auch Bühner (2006) äußert eine Präferenz der IRT gegenüber der KTT:

> Die Vorteile der [IRT] gegenüber der KTT sind weitgehend anerkannt [...]. Dazu gehören die spezifische Objektivität und erschöpfende Statistiken der Item- und Personenkennwerte (Bühner, 2006, S. 301)

Diese Aussage ist auch in Steyer und Eid (2001) widerlegt. Lewis (2007) zeigt sogar die „Äquivalenz“ des Modells tau-äquivalenter Messungen mit homoskedastischen normalverteilten Residualvarianzen und des Rasch-Modells. In beiden Modellen existieren suffiziente Statistiken, man kann eine CML-Schätzung ableiten und es gilt daher „spezifische Objektivität“. Es ist zu betonen, dass keine Normalverteilungsannahme des Traits θ in beiden Modellen notwendig ist.

Skalenniveau

Die Begrifflichkeit der Skalenniveaus von Daten ist ebenso mit vielfältigen Mythen behaftet (Niederée & Mausfeld, 1996a; Niederée & Mausfeld, 1996b). Es wird häufig behauptet, dass mit dem Rasch-Modell Messungen auf Intervallskalenniveau generiert werden könnten (Strobl, 2010, S. 24; Bühner, 2006, S. 300; siehe Fischer, 2007 für Details). Wir beleuchten im Folgenden diese Behauptung etwas näher.

In der repräsentativen Messtheorie (Pfanzagl, 1968; siehe auch Steyer & Eid, 2001) geht man von einem empirischen Relativ von einer Objektmenge und darauf definierten algebraischen Relationen (in einem Axiomensystem) aus. Das empirische Relativ wird auf ein numerisches Relativ auf der Menge reeller Zahlen und darauf definierten Relationen abgebildet. In diesem Konzept existieren keine Wahrscheinlichkeiten, sondern nur feste Objekte. Bei der Durchführung eines Tests ist aus meiner Sicht zunächst unklar, „was“ im Sinne der repräsentativen Messtheorie genau gemessen wird. Sind die Messobjekte Personen, Personengruppen oder Äquivalenzklassen von Personen, „misst man“ Verteilungen oder betrachtet man nur Eigenschaften des Messobjektes des Tests in Interaktion mit der Gesamtheit der Personen selbst? Es ist erstaunlich, dass die „nach der Messung“ eingesetzten statistischen Verfahren auf Wahrscheinlichkeiten beruhen, diese aber in der „gewöhnlichen“ Messtheorie nicht auftreten (siehe aber Heyer & Niederee, 1989, 1992 oder Rossi, 2006 für in den Sozialwissenschaften anwendbare Ausnahmen). Aufgrund der definitorischen und konzeptuellen Unverbundenheit von Messtheorie und statistischer Verfahren ist es nicht verwunderlich, dass die Messtheorie in den Sozialwissenschaften (aber vor allem auch den Naturwissenschaften, die in der Psychometrie als Ideal dient bzw. dienen sollte) praktisch keine Rolle spielt (vgl. Frigerio, Giordani & Mari, 2010).

Die repräsentative Messtheorie ist also nur auf einer Menge von Objekten definiert, die wir im Folgenden mit Personen oder Personengruppen assoziieren. Herleitungen für das Skalenniveau des Rasch-Modells setzen aber bereits an Axiomen auf der Ebene der Items an und „zeigen“ damit, dass Personenfähigkeiten und Itemschwierigkeiten auf einer Intervallskala abgebildet werden können (Fischer, 2007). Diese Ableitungen ergeben sich zunächst nur aus mathematischen Theoremen, die die Theorie der Funktionalgleichungen als Basis verwendet. Das Rasch-Modell wird dabei als Analog des *additive conjoint measurement* (ACM) gesehen (Perline, Wright & Wainer, 1979). Der Unterschied im Rasch-Modell zum eigentlichen ACM ist, dass die „Messungen“ im ACM modellimplizierte Wahrscheinlichkeiten des Rasch-Modells sind, also keine generischen Messungen, sondern nur Prädiktionen (van der Linden, 1994). Ob die Relationen zwischen den „gemessenen“ Wahrscheinlichkeiten tatsächlich den Bezug auf die Menge der Personen des empirischen Relativs erlauben, wird kontrovers diskutiert (Ballou, 2009; Borsboom & Zand Scholten, 2008; Kyngdon, 2008; van der Linden, 1994). Ich stimme Borsboom und Zand Scholten (2008) vorsichtig zu, dass die Prüfung von axiomatischen Eigenschaften für die Ableitung einer „Qualität“ eine gewisse Bedeutung besitzt. Es muss aber konstatiert werden, dass das Rasch-Modell in diesem Kontext bei der Ableitung des Skalenniveaus keine Ausnahmestellung besitzt. Wendet man anstatt des additiven conjoint measurement ein polynomiales conjoint measurement an, so „erhält“ man auch mit dem 2PL-Modell intervallskalierte Messwerte (Ballou, 2009; Kyngdon, 2011). Es ist überraschend, dass dennoch auch in aktuellen Arbeiten die Bedeutung des Rasch-Modells im Hinblick auf das Skalenniveau noch immer herausgehoben wird (Briggs, 2013; Domingue, 2014).

Ich nehme eher den Standpunkt ein, dass entweder das Skalenniveau einer Variablen ein (empirisch meistens nicht zugängliches) Postulat darstellt oder das Skalenniveau nur im Hinblick bedeutsamer statistischer Aussagen gewählt werden sollte (Niederée & Mausfeld, 1996a).

Zusammenfassend sehe ich durch die Verwendung des Rasch-Modells nicht die Möglichkeit „(spezifisch) objektivere Messungen“ als mit anderen IRT-Modellen abzubilden, die zusätzlich noch ein höheres Skalenniveau besitzen. Eher sehe ich das Rasch-Modell als Möglichkeit an, theoriegetrieben Items im Hinblick auf den Trait θ gleich zu gewichten (siehe auch Robitzsch et al., 2015), weshalb das Rasch-Modell für uns eine gewisse Nähe zu einem Item Sampling der klassischen Testtheorie oder der Generalisierbarkeitstheorie besitzt (Brennan, 2011; siehe auch Stenner, Burdick & Stone, 2008).

1.5 Mehrdimensionale IRT-Modelle

In vielen Kompetenztests oder Leistungstests sollen mehrere Kompetenzen zugleich gemessen werden. Statt einer Fähigkeit θ in eindimensionalen IRT-Modellen, wird in mehrdimensionalen IRT-Modellen (Reckase, 2009) eine mehrdimensionale Fähigkeit (d.h. ein Fähigkeitsvektor) $\boldsymbol{\theta}$ definiert. Wie in Faktorenanalysen unterscheidet man explorative und konfirmatorische mehrdimensionale IRT-Modelle (MIRT).

Wenn in konfirmatorischen MIRT-Modellen jedes Item auf genau einer Dimension lädt, so spricht man von *between item dimensionality*, ansonsten von *within item dimensionality* (Reckase, 2009; siehe auch Hartig, 2008). In einem *kompensatorischen 2PL-MIRT*-Modell

ist die Modellgleichung mit D Dimensionen wie folgt definiert (Reckase, 2009)

$$P(X_{pi}=1) = \Psi\left(\sum_{d=1}^{D} a_{id} q_{id} \theta_{pd} + d_i\right) \tag{1.26}$$

Die Q-Matrix $Q = (q_{id})$ mit dichotomen Einträgen gibt dabei an, welche Items auf welchen Dimensionen laden sollen[3]. Einige der Itemladungen a_{id} werden in konfirmatorischen Modellen praktisch auf Null fixiert. Der Item Intercept d_i entspricht der negativen Itemschwierigkeit b_i im eindimensionalen IRT-Modell. Das Modell (1.26) nimmt an, dass bei Vorliegen von within item dimensionality eine geringe Ausprägung in einer Dimension durch eine große Ausprägung in einer anderen Dimension im Hinblick auf die Lösungswahrscheinlichkeit kompensiert werden kann. Diese Annahme scheint in Anwendungen vielleicht nicht immer plausibel, weshalb dann das *nichtkompensatorische 2PL-MIRT*-Modell Einsatz findet (Reckase, 2009; siehe auch Bolt & Lall, 2003; Babcock, 2011; Wang & Nydick, 2015)

$$P(X_{pi}=1) = \prod_{d=1}^{D} \left[\Psi\left(a_{id}(\theta_{pd} - b_{id})\right)\right]^{q_{id}} \tag{1.27}$$

Während im kompensatorischen Modell (1.26) die Fähigkeiten $\boldsymbol{\theta}$ additiv in die gesamte Lösungswahrscheinlichkeit eingehen, gehen im nichtkompensatorischen Modell die Wahrscheinlichkeiten der „Lösung des Items auf den einzelnen Dimensionen“ multiplikativ eingehen. Dies ändert auch die Bedeutung der als mehrdimensional definierten Fähigkeit. Es wurden auch ein partiell kompensatorisches Modell vorgeschlagen, die für jedes Item einen weiteren Parameter mit Werten zwischen 0 und 1 beinhaltet, der einen Kompromiss aus vollständiger Kompensierbarkeit und Nichtkompensierbarkeit darstellt (Spray, Davey, Reckase, Ackerman & Carlson, 1990)[4].

Häufig wählt man eine multivariate Normalverteilung als Verteilungsannahme für den mehrdimensionalen Trait $\boldsymbol{\theta}$. Es können allerdings auch diskrete Verteilungen gewählt werden, die die Normalverteilung recht gut approximieren (Haberman, von Davier & Lee, 2008). Eine Verbindung parametrischer Item-Response-Funktionen und einer Repräsentation der Trait-Verteilung von $\boldsymbol{\theta}$ als Cluster stellen die *multidimensional latent class IRT models* dar (Bartolucci, 2007; Bartolucci, Montanari & Pandolfi, 2012). Kognitivdiagnostische Modelle als spezielle MIRT-Modelle spezifizieren diskrete latente Fähigkeiten (Rupp & Templin, 2008). Alternativen zur Normalverteilung und alternative Linkfunktionen werden in *Copula-Modellen* diskutiert (Brechmann & Joe, 2014; Krupskii & Joe, 2013; Nikoloulopoulos & Joe, 2015; siehe Joe, 2015 für einen Überblick).

Eine Version des Binomialmodells für mehrere Dimensionen ist im Kontext kognitivdiagnostischer Modelle von Hong, Wang, Lim und Douglas (2015) vorgeschlagen worden. Das vorgeschlagene *continuous conjunctive model* (CCM) wird wie folgt definiert

$$P(X_{pi}=1) = \prod_{d=1}^{D} \xi_{pd}^{q_{id}} \qquad \text{mit } 0 \leq \xi_{pd} \leq 1 \tag{1.28}$$

[3]Siehe DeCarlo (2012) für einen Ansatz, der auch eine mögliche Unreliabilität der Einträge in der Q-Matrix berücksichtigt.

[4]Eine Implementierung des Modells ist in der Funktion `smirt` im R-Paket `sirt` (Robitzsch, 2015) vorgenommen.

Die Fähigkeiten ξ_{pd} kann man als Anteile verstehen, mit denen die Dimension d für ein jeweiliges Item i „gelöst“ wird[5]. Für eine Beschreibung der Verteilung über alle Personen kann eine multivariate Normalverteilung für $\boldsymbol{\theta} = (\theta_d)_d$ geschätzt werden und die Anteile ξ_d ergeben sich gemäß $\xi_d = \Phi(\theta_d)$.

1.6 IRT-Modelle, log-lineare Modelle und Ising-Modell

Ein- und mehrdimensionale Item-Response-Modelle repräsentieren eine hochdimensionale Kontingenztafel $P(\boldsymbol{X})$ mit 2^I verschiedenen Wahrscheinlichkeiten (sofern alle Item Response Pattern beobachtet werden) durch Einführung einer ein- oder mehrdimensionalen latenten Variablen. Man kann also formal mit einer Verteilungsfunktion F der Fähigkeit θ und unter der Annahme der lokalen stochastischen Unabhängigkeit schreiben (vgl. Holland, 1990a)

$$P(\boldsymbol{X} = \boldsymbol{x}) = P(\boldsymbol{x}) = \int \left(\prod_{i=1}^{I} P(X_i = x_i|\theta) \right) \mathrm{d}F(\theta) \tag{1.29}$$

wobei sowohl die Item-Response-Funktionen $P(X_i = x_i|\theta)$ als auch die Verteilung F parametrisiert sind. Die Item Response Pattern $\boldsymbol{X} = (X_i)$ werden dann durch die Einführung der latenten Variablen θ parametrisiert und die Zusammenhänge zwischen den Items daher durch θ (bzw. dessen Verteilung F) beschrieben. Diese Sichtweise entspricht der mittleren Grafik in Abbildung 1.1.

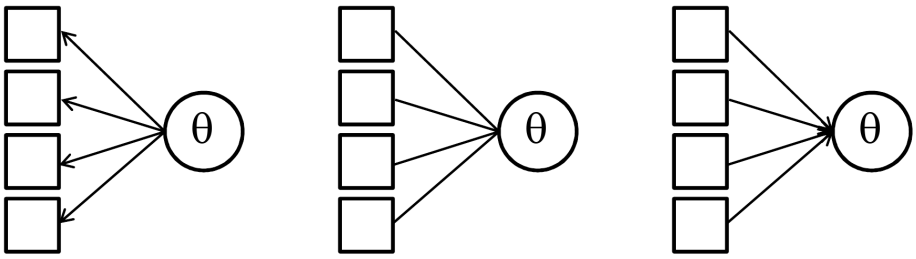

Abbildung 1.1: *Links: Reflektives Modell, Mitte: Korrelatives Modell, Rechts: Formatives Modell*

Häufig interpretiert man die latente Variable θ allerdings in einem kausalen Sinne, in dem Variation in Fähigkeiten eine Variation in den Item Responses X_i induziert (Borsboom, 2005). Diese Perspektive wird in der linken Grafik von Abbildung 1.1 dargestellt. Die Pfeile von der latenten Variablen auf die Items zeigen eine „Wirkrichtung“ an.

Holland (1990b) zeigt in der sog. *Dutch identity*, dass sich viele IRT-Modelle als log-lineares Modell zweiter Ordnung repräsentieren lassen, d.h. sich in der Form

$$\log P(\boldsymbol{x}) = \alpha + \sum_i \beta_i x_i + \sum_i \sum_j \omega_{ij} x_i x_j \tag{1.30}$$

[5]Man kann zeigen, dass sich mit diesem Modell eine reparametrisierte Version des nicht identifizierbaren NIDA-Modells (Junker & Sijtsma, 2001) schätzen lässt.

schreiben lassen (siehe auch Anderson & Vermunt, 2000). Dies ist demnach nur eine Repräsentation der Verteilung $P(\boldsymbol{X})$ von $\boldsymbol{X}$ und greift nicht auf eine latente Variable zurück. Holland (1990b) zeigt, dass bei Gültigkeit eines mehrdimensionalen IRT-Modells mit Itemladungen a_{id} in die Zusammenhangsparameter ω_{ij} in (1.30) überführen lassen. Für das Rasch-Modell ergibt sich folgende einfache Form des log-linearen Modells (Holland, 1990b, S. 9)

$$\log P(\boldsymbol{x}) = \alpha + \sum_i \beta_i x_i + \sum_j \gamma_j \mathbf{1}_{\{\sum_i x_i = j\}} \tag{1.31}$$

Für jeden der Summenscores $j = 0, 1, \ldots, I$ gibt es in (1.31) einen eigenen γ-Koeffizienten.

Die Darstellung der IRT-Modelle als log-lineares Modell hat bei vielen Items eher konzeptuelle als computationale Vorteile, denn log-lineare Modelle mit vielen Item Response Pattern sind schwierig anzupassen. Anderson und Yu (2007) zeigen jedoch für das Rasch-Modell, dass zur Gewinnung von Itemparametern kein log-lineares Modell, sondern nur eine Folge logistischer Regressionen angepasst werden muss, wobei der Item Response X_i die abhängige Variable und der Restscore $X_+^{(-i)} = \sum_j X_j - X_i$ als unabhängige Variable verwendet wird (Anderson, Li & Vermunt, 2007). Diese Berechnungsweise zeigt, dass die bedingte Verteilung eines bestimmten Item durch eine Zusammenfassung aller anderen Items (dem Restscore) vorhergesagt werden kann. In einem metaphorischen Sinne könnte man also sagen, dass der „Test sich selbst definiert“ und die latente Variable θ im Rasch-Modell aus den Item Responses selbst definiert wird, wie dies in der rechten Grafik von Abbildung 1.1 dargestellt wird (siehe Anderson & Yu, 2007, S. 10).

In neuerer Literatur werden sog. Netzwerkmodelle (*network models*) vorgeschlagen, um multivariate Zusammenhänge zwischen manifesten Item Responses zu untersuchen (Cramer et al., 2012; Schmittmann et al., 2013). Häufig stellt man diese Netzwerkmodelle in Form eines gewichteten Graphen dar, wobei Variablen die Ecken und die gewichteten Kanten die Zusammenhänge der Items darstellen. Existiert kein Zusammenhang zwischen bestimmten Items, so wird keine Kante im Graphen dargestellt. Netzwerkmodelle verzichten auf latente Variablen und erklären Zusammenhänge zwischen Variablen dadurch, dass die Änderung in der Ausprägung einzelner Items die Ausprägungen anderer Items verursacht (Epskamp, Maris, Waldorp & Borsboom, 2015). Dabei wird ein Vorteil darin gesehen, nicht mehr die latente Variable als kausal für die Item Responses interpretieren zu müssen (Epskamp et al., 2015).

Für die Schätzung von Netzwerken mit dichotomen Daten wurde das *Ising-Modell* vorgeschlagen (van Borkulo et al., 2014; Epskamp et al., 2015; siehe auch Bühlmann & Van De Geer, 2011). Das Ising-Modell entspricht formal einem log-linearen Modell zweiter Ordnung (wie in (1.30) definiert). Die Schätzung dieses Modells (insbesondere die ω_{ij}-Parameter) erfolgt mit Hilfe von logistischen Regressionen, wobei jedes Item jeweils einmal die abhängige Variable darstellt und alle restlichen Items die unabhängige Variablen. Zur Vermeidung des Overfitting werden regularisierte logistische Regressionen verwendet (van Borkulo et al., 2014; Epskamp et al., 2015).

Epskamp et al. (2015) zeigen, dass das Ising-Modell äquivalent zu einem mehrdimensionalen IRT-Modell einer geeigneten Dimension ist. Daher kann man argumentieren, dass sowohl die Perspektive latenter Variablen mit reflexiven Messmodellen als auch die Netzwerk-Perspektive sinnvolle Interpretationen erlauben. In computationaler Hinsicht ist das Ising-Modell bei Verwendung einer regularisierten Schätzung dem log-linearen

Modell aufgrund der Spezifikation als Folge regularisierter logistischer Regressionsmodelle vorzuziehen. Netzwerk-Modelle können als parametrisch sparsame Repräsentation einer hochdimensionalen Kontingenztafel dienen. Erfolgreiche empirische Anwendungen müssen allerdings erst als Vorbild dafür dienen, um Netzwerkmodelle eine substanzielle Bedeutung neben den dominierenden reflexiven Messmodellen mit latenten Variablen zu geben.

1.7 Restringierte Latent-Class-Modelle

In diesem Abschnitt argumentieren wir, dass *Latent-Class-Modelle* (LCA; Formann, 1984) eine recht allgemeine Modellklasse darstellen, in die sich die bisher diskutierten Item-Response-Modelle gut einordnen lassen. Die Grundidee besteht darin, dass man Zusammenhänge zwischen Items durch wenige zunächst als kategorial angenommene Klassen beschreiben. Das Prinzip ist in Abbildung 1.2 dargestellt.

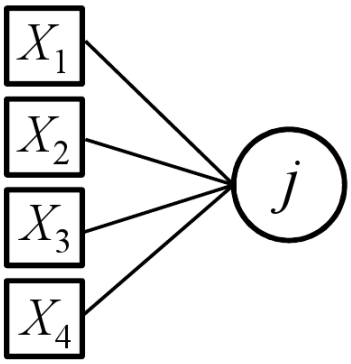

Abbildung 1.2: *Latent-Class-Modell*

Für vier Items X_i $(i = 1, 2, 3, 4)$ sind die Zusammenhänge durch Zugehörigkeit zu einer Klasse j erklärbar. Ein LCA-Modell ist dann durch die Annahme der lokalen stochastischen Unabhängigkeit definiert

$$P(\boldsymbol{X} = \boldsymbol{x}|C = j) = \prod_{i=1}^{I} P(X_i = x_i|C = j) = \prod_{i=1}^{I} \left(p_{i|j}^{x_i} (1 - p_{i|j})^{1-x_i} \right) \tag{1.32}$$

Die Items sind also bedingt auf die Klasse j unabhängig. Die Gleichung (1.32) ist äquivalent zu unkorrelierten Residuen in Faktormodellen. In einem saturierten LCA-Modell sind alle Item-Response-Funktionen $p_{i|j} = P(X_i = 1|C = j)$ zu schätzen. Werden also K Klassen angenommen, so sind für I Items genau $I \cdot K$ Itemparameter zu schätzen. Außerdem wird die Verteilung der Klassenvariablen C bestimmt, d.h. die Wahrscheinlichkeiten $w_j = P(C = j)$. Bei K Klassen sind demzufolge $K - 1$ Wahrscheinlichkeiten frei zu schätzen. Die Annahme in Latent-Class-Modellen ist, dass jede Person (oder Personengruppe) genau einer Klasse j zugeordnet werden kann. Damit ergibt sich als gesamte

Wahrscheinlichkeit für das Item Response Pattern $\boldsymbol{x}$ (vgl. Erosheva, 2006)

$$P(\boldsymbol{X} = \boldsymbol{x}) = \sum_j P(\boldsymbol{X} = \boldsymbol{x}|C = j)P(C = j) = \sum_j \prod_i \left(p_{i|j}^{x_i}(1 - p_{i|j})^{1-x_i} \right) w_j \tag{1.33}$$

In Anwendungen besteht das Problem, dass die Anzahl K der latenten Klassen bestimmt werden muss (siehe Formann, 1984). Häufig geschieht dies auf Basis von Informationskriterien oder unter Beurteilung des Modellfits. Ein LCA-Modell mit theoretisch unendlich vielen latenten Klassen ist von Dunson und Xing (2009) vorgeschlagen wurden (*Dirichlet process mixture of products of multinomial distributions model*; *DPMPM*; siehe auch Si & Reiter, 2013), wobei dabei die praktisch notwendige Klassenanzahl während der Modellschätzung ermittelt wird, in dem eine hierarchische Verteilung für die Klassenwahrscheinlichkeiten spezifiziert wird. In diesem Modell wird jedoch angenommen, dass die Wahrscheinlichkeiten der Klassen mit einem höheren Index gegen Null konvergieren, d.h. man nimmt $w_1 \geq w_2 \geq w_3 \geq \ldots$ (im stochastischen Sinn) an. Gewöhnliche Latent-Class-Modelle sind nur bis auf beliebige Permutationen der Klassenvariablen eindeutig (d.h. die Klassen können umnummeriert werden). Das DPMPM besitzt jedoch dieses Problem nicht und könnte daher ein robusteres Modell darstellen.

Wir bemerken, dass die Anpassung des LCA-Modells (1.33) für jedes Item-Response-Pattern $\boldsymbol{x}_p$ einer Person p Posteriorwahrscheinlichkeiten $P(C = j|\boldsymbol{X} = \boldsymbol{x}_p)$ berechnet werden können, die die Wahrscheinlichkeit der Zuordnung der Person p zu Klasse j beschreibt. Dabei gilt

$$P(C = j|\boldsymbol{X} = \boldsymbol{x}_p) \propto P(\boldsymbol{X} = \boldsymbol{x}_p|C = j) \cdot P(C = j) \tag{1.34}$$

Dabei geht die Priorwahrscheinlichkeit $P(C = j)$ und die Likelihood $P(\boldsymbol{X} = \boldsymbol{x}_p|C = j)$ in die Posteriorwahrscheinlichkeit ein. Es ist wichtig zu betonen, dass „nur“ Unreliabilität dazu führt, dass die Zuordnung einer Person zu einer Klasse nicht eindeutig vorgenommen werden kann. Das statistische Modell nimmt hinwegen auf Populationsebene an, dass jede Person genau einer Klasse zugeordnet werden kann und das Lösungsverhalten der Person auf den Items genau durch diese Klasse charakterisiert wird.

So genannte *restringierte Latent-Class-Modelle* (restricted latent class models; RLCA) führen für die Item-Response-Funktion $p_{i|j}$ und die Klassenwahrscheinlichkeiten w_j Restriktionen ein (Formann, 1984). Ähnlich zum linear-logistischen Testmodell (LLTM; Fischer, 1973) schlägt Formann (1985) ein linear-logistisches Latent-Class-Modell vor. In diesem werden die Item-Response-Wahrscheinlichkeiten jedes Items i und jeder Klasse j der Logittransformation unterworfen und eine Linearkombination angenommen. Dabei gilt (vgl. Formann, 1985)

$$\operatorname{logit} p_{i|j} = \beta_{ij} = \sum_{v=1}^{V_\beta} q_{\beta,ijv} \lambda_{\beta,v} \tag{1.35}$$

Dabei werden die Wahrscheinlichkeiten $p_{i|j}$ durch den Parameter β_{ij}, die sich als eine Linearkombination von V_β Basisparametern $\lambda_{\beta,v}$ schreiben lassen, wobei die Gewichte $q_{\beta,ijv}$ vorgegeben werden. Für den Logit der Klassenwahrscheinlichkeiten w_j nimmt Formann (1985) ebenso eine Linearkombination von Basisparametern an. Nach unseren Erfahrungen ist die Verwendung der log-Transformation für Wahrscheinlichkeiten (siehe Xu & von

Davier, 2008) computational stabiler. Dabei setzt man

$$\log w_j = \sum_{v=1}^{V_w} q_{w,v} \lambda_{w,v} \tag{1.36}$$

an, wobei V_w Basisparameter geschätzt werden. Man kann sich (1.36) als „Glättung“ der Klassenwahrscheinlichkeiten w_j vorstellen.

Es ist offensichtlich, dass das unrestringierte Latent-Class-Modell (1.32) ein Spezialfall des restringierten Latent-Class-Modells darstellt. Formann (1984) zeigt, dass sich ein- und mehrdimensionale IRT-Modelle als restringiertes Latent-Class-Modell darstellen lassen (vgl. auch Heinen, 1993). Eine Normalverteilung lässt sich näherungsweise durch Wahrscheinlichkeiten in einer RLCA darstellen, in dem man w_j proportional zu $\phi(\theta_j)$ wählt, wobei θ_j vorgegebene Stützstellen der stetigen θ-Verteilung und ϕ die Dichtefunktion der Normalverteilung bezeichnet (siehe auch ein ähnliches Vorgehen im *general diagnostic model*; von Davier, 2008). Im Sinne der RLCA könnte man die Aussage in Gifi (1990) verstehen, dass stetige Verteilungen Approximationen von diskreten Verteilungen darstellen. Die RLCA „regularisiert“ demzufolge diskrete Klassenwahrscheinlichkeiten w_j durch die Annahme einer stetigen Verteilungsfamilie.

Für das Rasch-Modell wollen wir das Vorgehen im Rahmen der RLCA illustrieren. Die Wahrscheinlichkeiten w_j können auf eine diskret repräsentierte Normalverteilung fixiert werden (man wähle z.B. 21 äquidistante Stützstellen von -6 bis 6). Die zu schätzenden Parameter σ und b_i im Rasch-Modell lassen sich dann mittels der Item-Response-Funktionen in der Form

$$\operatorname{logit} p_{i|j} = \beta_{ij} = \theta_j \cdot \sigma + (-1) \cdot b_i \tag{1.37}$$

schreiben. Der zu schätzende Basisparameter $\boldsymbol{\lambda}_\beta$ ist demzufolge der Vektor $(\sigma, b_1, b_2, \ldots, b_I)$ und die vorgegebenen Gewichte $q_{\beta,ijv}$ sind mit θ_j und -1 entsprechend der Gleichung (1.37) zu wählen. Die Verallgemeinerung des Prinzips auf die Schätzung im 2PL-Modell scheint offensichtlich.

Die RLCA erlaubt beispielsweise die Spezifikation mehrdimensionaler kompensatorischer IRT-Modelle, located latent class models und Mischverteilungs-IRT-Modellen wie des Mixed Rasch Models (siehe Formann, 1985, 2007; Formann & Kohlmann, 1998, 2002). Auch kognitiv-diagnostische Modelle sind als Spezialfall der RLCA-Modelle anzusehen (vgl. von Davier, 2009).

Das linear-logistische Latent-Class-Modell für dichotome Daten (Formann, 1985) ist auf polytome Daten erweitert wurden (Formann, 1992; siehe auch schon Formann, 1984), so dass beispielsweise das Generalized Partial Credit Modell damit geschätzt werden kann. Die polytome Version des RLCA ist im R-Paket CDM (Robitzsch, Kiefer, George & Ünlü, 2015) in der Funktion `slca` implementiert.

Formann und Kohlmann (1998, 2002) verallgemeinern das linear-logistische Latent-Class-Modell mit dem Intercept-Parameter β_{ij} im sog. *structured latent class model* (SLCA) um einen Trennschärfeparameter α_{ij} und einen Rateparameter γ_{ij}, so dass man insgesamt vom Modell

$$p_{i|j} = \gamma_{ij} + (1 - \gamma_{ij}) \cdot \Psi(\alpha_{ij} \cdot \beta_{ij}) \tag{1.38}$$

Diese Parameter repräsentiert man wiederum durch Basisparameter

$$\beta_{ij} = \sum_{v=1}^{V_\beta} q_{\beta,ijv}\lambda_{\beta,v} \tag{1.35}$$

$$\log\, \alpha_{ij} = \sum_{v=1}^{V_\alpha} q_{\alpha,ijv}\lambda_{\alpha,v} \tag{1.39}$$

$$\operatorname{logit} \gamma_{ij} = \sum_{v=1}^{V_\gamma} q_{\gamma,ijv}\lambda_{\gamma,v} \tag{1.40}$$

Damit können beispielsweise klassenspezifische (d.h. personenspezifische) Rateparameter geschätzt werden (siehe Raiche et al., 2013), auch wenn die Identifikation dieser Parameter schwierig sein kann (Formann & Kohlmann, 2002). In diesem SLCA-Modell können auch simultan klassen- und itemspezifische Trennschärfen angenommen werden, womit Person Misfit abgebildet werden kann (vgl. Ferrando, 2007; Strandmark & Linn, 1987).

Für die Modellierung mehrdimensionaler Kompetenzen argumentiert Wilson (2013), dass man Hypothesen über diese Kompetenzen einfacher auf Basis ordinaler latenter Variablen anstelle stetiger Variablen testen könne (siehe auch Minnamaier, 2002). Die ordinalen latenten Variablen können dann als „Kompetenzstufen“ interpretiert werden. Insbesondere können Hypothesen über die mehrdimensionale Verteilung der latenten Variablen getestet werden. Beispielsweise kann der Fall ausgeschlossen werden, dass sich eine Person in Kompetenzbereich A auf Stufe 5, in Kompetenzbereich B jedoch auf Stufe 1 befindet (Wilson, 2012). Die hier diskutierten Modelle sind mehrdimensionale Latent-Class-Modelle mit geordneten Klassen (Hoijtink & Molenaar, 1997; siehe auch Shojima, 2007).

Zusammenfassend sehen wir die restringierten LCA-Modelle zentral in der IRT und als so flexibel an, dass praktisch alle relevanten psychometrischen Modelle für diskrete Item Responses damit geschätzt bzw. repräsentiert werden können. Leider scheint die von Formann (1984) entwickelte und von Formann und Kohlmann (1998) erweiterte Modellklasse nicht den Stellenwert in der psychometrischen Literatur zu genießen, den sie eigentlich aufgrund der Relevanz besitzen sollte.

1.8 Unscharfe latente Variablen und unscharfe Item Responses

Im letzten Abschnitt erweitern wir unsere Überlegungen zu IRT-Modellen auf *unscharfe Daten* (auch *Fuzzy-Daten*; z.B. Viertl, 2006). Dabei werden Beobachtungen (bzw. Messungen) nicht mit einer eindeutigen Zahl assoziiert, sondern nur durch einen Bereich möglicher Werte wird die „Unsicherheit“ ausgedrückt. Wir gehen auf unscharfe latente Variablen und unscharfe Item Responses ein.

Unscharfe latente Variablen

Wir führen psychometrische Modelle mit unscharfen latenten Variablen als Verallgemeinerung von Latent-Class-Modellen ein (Erosheva, Fienberg & Junker, 2002). In sog. *mixed membership models* (Erosheva, 2005; Erosheva, 2006; Erosheva, Fienberg & Joutard, 2007; siehe auch Gruhl & Erosheva, 2015 oder Galyart, 2015 für einen Überblick) wird angenommen, dass eine Person graduell Klassen zugehören kann. Beispielsweise könnte in einem Modell mit drei Klassen eine bestimmte Person mit dem „Anteil“ .60 zu Klasse 1, dem Anteil .35 zu Klasse 2 und dem Anteil .05 zu Klasse 3 zugeordnet sein. Eine alternative Interpretation besteht darin, dass eine Person für jedes Item in eine andere Klasse „wechseln“ kann. Im Beispiel würde dies heißen, dass eine Person für 60% Items nach Klasse 1, für 35% für Items nach Klasse 2 und für 5% nach Klasse 3 beantworten würde. Das gewöhnliche Latent-Class-Modell ergibt sich, wenn alle Personen mit einem Anteil von 1 genau einer Klasse zugeordnet werden.

Mixed membership models könnte beispielsweise bei mentalen Rotationsaufgaben relevant sein, wenn Personen Items mit verschiedenen Strategien lösen können. Jede Strategie soll dabei durch eine latente Klasse repräsentiert sein. Wenn ein mixed membership für eine Person vorliegt, so bedeutet das, dass diese Person itemspezifisch eine andere Strategie (also latente Klasse) auswählt. In mixed mebership models ist also ein intraindividuelles „Springen“ zwischen latenten Klassen zugelassen.

Diese Konzeption legt nahe, dass in Modellen mit K Klassen und einer unscharfen latenten Variablen jeder Person ein Vektor von *membership scores* $\boldsymbol{g} = (g_1, \ldots, g_K)$ mit $\sum_{k=1}^{K} g_k = 1$ zugeordnet wird. Als Verteilung für diesen Vektor mit $K-1$ freien Parametern wird häufig die Dirichlet-Verteilung (Erosheva et al., 2007) oder eine $K-1$-dimensionale Normalverteilung gewählt (Asparouhov & Muthen, 2008; Blei & Lafferty, 2007; Gruhl & Erosheva, 2015), die den Vektor der membership scores repräsentieren. Ist der $K-1$-dimensionale Vektor $\boldsymbol{\xi} = (\xi_k)_k$ multivariat normalverteilt, so definiert man für $k = 1, \ldots, K-1$ den membership score $g_k := \exp(\xi_k) / \left(1 + \sum_{k=1}^{K-1} \exp(\xi_k)\right)$ sowie $g_K := 1 / \left(1 + \sum_{k=1}^{K-1} \exp(\xi_k)\right)$ für die Klasse K. Bei einer Repräsentation des mixed membership models durch die multivariate Normalverteilung werden demzufolge $K-1$ Mittelwerte, $K-1$ Varianzen und $(K-1) \cdot (K-2)/2$ Kovarianzen geschätzt. Für ein mixed membership model mit $K = 3$ Klassen ergeben sich dann insgesamt $2 + 2 + 1 = 5$ zu schätzende Parameter, während im latent class model nur 2 Parameter (2 Wahrscheinlichkeiten) zu schätzen sind. Das latente-Klassen-Modell ergibt sich aus dem mixed membership model, in dem alle Kovarianzen auf 0 und alle Varianzen auf einen großen Wert (z.B. 1000) fixiert werden. Nur die Mittelwerte der multivariaten Normalverteilung sind dann noch frei zu schätzen, die die Klassenwahrscheinlichkeiten in der LCA darstellen.

Im mixed membership model werden zunächst Item-Response-Funktionen definiert, die für Personen gelten, die genau einer Klasse zugeordnet sind. Dabei ist $P(X_i = 1 | g_j = 1) = p_{i|j}$ die Wahrscheinlichkeit das Item i zu lösen, wenn sich eine Person *immer* in Klasse j befindet. Für Klassenzugehörigkeiten zwischen 0 und 1 wird die Item-Response-Funktion unter der Annahme von *mixed membership* definiert als

$$P(X_i = 1 | g_1, \ldots, g_K) = p_{i|\boldsymbol{g}} = \sum_{j=1}^{K} g_j P(X_i = 1 | g_j = 1) = \sum_{j=1}^{K} g_j p_{i|j} \qquad (1.41)$$

Ein Item-Response-Pattern $\boldsymbol{x}$ wird unter Annahme einer Verteilung F für $\boldsymbol{g}$ dann unter Annahme der lokalen stochastischen Unabhängigkeit wie folgt repräsentiert

$$P(\boldsymbol{X} = \boldsymbol{x}) = \int \left(\prod_i p_{i|\boldsymbol{g}}^{x_i} (1 - p_{i|\boldsymbol{g}})^{1-x_i} \right) \mathrm{d}F(\boldsymbol{g}) \tag{1.42}$$

Die Verteilungsparameter für F und die *extremalen* Item-Response-Wahrscheinlichkeiten $p_{i|j}$ können mit MCMC-Verfahren (Erosheva et al., 2007) oder einem EM-Algorithmus (Bingham, Kabán & Fortelius, 2009) geschätzt werden. Asparouhov und Muthen (2008) zeigen, dass man mixed membership models als ein Multilevel Latent Class Model (Vermunt, 2003; Vermunt & Magidson, 2005) spezifizieren kann. Auf Level 2 werden dabei Personen repräsentiert. Auf Level 1 wird für für jede Person und jedes Item eine latente Klasse „ausgewählt“, wobei sich die Item-Response-Wahrscheinlichkeit dann durch $p_{i|j}$ ergibt, falls sich für dieses Item i die Person in Klasse j befindet (siehe auch Erosheva, Fienberg & Lafferty, 2004).

Die membership scores $\boldsymbol{g}$ sollten nicht Posteriorwahrscheinlichkeiten individueller Personen in einem Latent-Class-Modell verwechselt werden. Auch im mixed membership model können Posteriorwahrscheinlichkeiten $P(\boldsymbol{G} = \boldsymbol{g}|\boldsymbol{X} = \boldsymbol{x}_p)$ berechnet werden können. Diese Größen stellen nun allerdings eine Verteilung für den mehrdimensionalen Vektor $\boldsymbol{g}$ der membership scores dar. Insofern ist das mixed membership model flexibler als LCA-Modelle.

Anstelle der Darstellung der Wahrscheinlichkeit $P(X_i = 1|\boldsymbol{g})$ als Summe verwendet man im Ansatz des *partial membership models* (Gruhl & Erosheva, 2015; Ghahramani, Mohamed & Heller, 2015) eine multiplikative Form

$$P(X_i = 1|g_1, \ldots, g_K) = p_{i|\boldsymbol{g}} = \prod_{j=1}^{K} p_{i|j}^{g_j} \tag{1.43}$$

Aufgrund der Multiplikativität in (1.44) können existierende EM-Algorithmen von gewöhnlichen (restringierten) Latent-Class-Modellen leicht angepasst werden, denn die Logarithmierung ergibt

$$\log p_{i|\boldsymbol{g}} = \sum_{j=1}^{K} g_j \cdot \log p_{i|j} \tag{1.44}$$

Die membership scores g_j gehen dann in EM-Algorithmen direkt in die expected counts ein.

Es sei angemerkt, dass die Ansatz unscharfer latenter Variablen mit membership scores $\boldsymbol{g}$ auf beliebige IRT-Modelle anwendbar ist. Dabei eignet sich insbesondere der Ansatz der restringierten Latent-Class-Modelle von Formann (1984), der „nur noch“ um Verteilungen für die membership scores ergänzt werden muss.

Nehmen wir abschließend an, dass ein eindimensionales IRT-Modell mit Item-Response-Funktionen $P(X_i = k|\theta)$ (für $k = 0, 1$) gelten möge. Überträgt man das Konzept unscharfer latenter Variablen auf stetige Variablen, so wird jede Person (oder Personengruppe) nicht mit einer Fähigkeit θ, sondern einer *membership function* $\boldsymbol{g} = g(\theta)$ assoziiert (wir

fordern $\int g(\theta)\mathrm{d}\theta = 1$). Die Wahrscheinlichkeit, dass eine Person ein Item i richtig löst, ist anstelle der Summation in (1.41) nun durch eine Integration zu erhalten

$$P(X_i = 1|\boldsymbol{g}) = p_{i|\boldsymbol{g}} = \int g(\theta)P(X_i = 1|\theta)\mathrm{d}\theta \tag{1.45}$$

Typischerweise wird man in Anwendungen die Zugehörigkeits-Funktion $\boldsymbol{g}$ mit einer vorgegebenen parametrischen Form schätzen. Für eine Person p könnte man annehmen, dass $g_p(\theta) = F(\theta; \theta_p, \alpha_p)$ erfüllt ist, wobei F eine vorgegebene parametrische Funktion in den Parametern θ_p (z.B. der Fähigkeit) und α_p (z.B. Person Misfit) bezeichnet. Damit wird das Problem der Schätzung einer Verteilung für einen unendlich-dimensionalen Vektor $\boldsymbol{g}$ (einem Funktionenraum) auf ein zweidimensionales IRT-Modell (mit einer Verteilung für (θ, α)) reduziert.

Unscharfe manifeste Variablen

Neben unscharfen latenten Variablen können die Item Responses selbst unscharf sein. Dies kann beispielsweise bei der Bewertung eines Items durch einen Rater auftreten, bei dem eine Begründung durch den Schüler verlangt wird, die Kategorie 1 (richtig) oder 0 (falsch) nicht eindeutig vergeben werden kann oder eine gewisse „Unsicherheit“ vorliegt. Dabei wird diese Unsicherheit im Allgemeinen nicht mit den Wahrscheinlichkeiten im IRT-Modell abgebildet, denn diese bezieht sich auf die Repräsentation von *scharfen Daten* (d.h. eindeutig durch 0 oder 1 bewertete Items) durch latente Variablen. Ein anderes Beispiel stellt die holistisch orientierte Beurteilung von Schreibaufgaben dar, bei der man nicht von einer Existenz einer „wahren Kategorie“ eines Items ausgehen kann.

Wenn wir wiederum nur den Fall dichotomer Items betrachten, so wird anstelle des dichotomen Item Response x_{pi} bei unscharfen Daten nun von membership scores (m_{pi0}, m_{pi1}) ausgegangen, wobei $m_{pi0} + m_{pi1} = 1$ gilt. Dabei bezeichnet m_{pi1} die Zugehörigkeit der Person p auf Item i zur Kategorie 1 oder die „Sicherheit“ für eine richtige Antwort. Man kann sich unter m_{pi1} auch eine Wahrscheinlichkeit vorstellen. Für scharfe Daten x_{pi} kann man mit der Item-Response-Funktion $P_i(\theta) = P(X_i = 1|\theta)$ die log-Likelihood für Person p und Item i wie folgt formulieren

$$\log L(x_{pi}|\theta_p) = x_{pi} \log\, P_i(\theta_p) + (1 - x_{pi}) \log\,(1 - P_i(\theta_p)) \tag{1.46}$$

Für unscharfe Item Responses (*Fuzzy Item Responses*) schlägt Denoeux (2013) eine Verallgemeinerung des EM-Algorithmus für die Schätzung vor. In diesem Ansatz verallgemeinert er die Likelihood (1.46) auf die membership scores $m_{pi0} = 1 - m_{pi1}$ und m_{pi1} durch

$$\log L(m_{pi1}|\theta_p) = \log\left\{m_{pi1}P_i(\theta_p) + (1 - m_{pi1})(1 - P_i(\theta_p))\right\} \tag{1.47}$$

Analog zu unscharfen latenten Variablen kann man also von einem mixed membership von Person p und Item i sprechen (siehe Gleichung (1.41); Gruhl & Erosheva, 2015). In der psychometrischen Literatur wurde jedoch von Lord (1974) ein Pseudo-Likelihood-Ansatz für unscharfe Item Responses im Kontext der Bewertung fehlender Multiple-Choice-Items vorgeschlagen. Dabei ist die Likelihood im Sinne der partial membership (vgl. Gleichung (1.44))

$$\log L(m_{pi1}|\theta_p) = m_{pi1} \log\, P_i(\theta_p) + (1 - m_{pi1}) \log\,(1 - P_i(\theta_p)) \tag{1.48}$$

definiert. In diesem Ansatz müssen existierende EM-Algorithmen für IRT-Modelle (nahezu) nicht modifiziert werden, da membership scores m_{pi1} direkt in die Berechnung der expected counts eingehen.

Zusammenfassend argumentieren wir, dass für manche psychometrische Fragestellungen eine Perspektive unscharfer Daten (bzw. mixed membership) fruchtbar sein könnte. Insbesondere könnte die restringierten Latent-Class-Modelle um unscharfe latente Variablen und unscharfe Item Responses erweitert werden. Dadurch könnte in Anwendungen ine große Anzahl von Klassen in gewöhnlichen Latent-Class-Analysen auf eine kleinere Anzahl von „Extremklassen“ in mixed membership models reduziert werden, die eine einfachere Interpretation erlauben. Mixed membership models erlangen immer dann Bedeutung, wenn Personen „Cluster“ (d.h. latente Klassen) nicht eindeutig zugeordnet werden können.

Kapitel 2

Fragestellungen der Arbeit

In diesem Kapitel werden die Fragestellungen der vorliegenden Arbeit eingeführt und ein grober Überblick zu jedem Kapitel gegeben.

2.1 Kapitel 3: Ausgewählte methodische Herausforderungen bei der Kalibrierung von Leistungstests

In Kapitel 3 werden ausgewählte methodische Fragestellungen bei der Kalibrierung von Leistungstests diskutiert. Dabei liegt der Fokus auf dem häufig in der Kalibrierung als Skalierungsmodell eingesetzten Rasch-Modell. Die Annahmen dieses Modells werden praktisch jedoch häufig verletzt sein. In Kapitel 3 werden daher einige Item-Response-Modelle als Erweiterungen des Rasch-Modells illustrativ vorgestellt und dessen Einsatz kritisch diskutiert.

Wie in Kapitel 1 näher erläutert wurde, besitzt das Rasch-Modelle gegenüber anderen Skalierungsmodellen einige hervorstechende Eigenschaften. Der Summenscore im Test ist eine suffiziente Statistik für die Personenfähigkeit, so dass damit alle Items gleich gewichtet in den Test eingehen. Außerdem sind die Item-Response-Funktionen parallel, es gilt daher die in Abschnitt 1.4 kritisch diskutierte spezifische Objektivität. In Abschnitt 3.2 widmen wir uns der Bedeutung der latenten Variablen im Rasch-Modell. Bereits in Abschnitt 1.6 wurde festgestellt, dass man anstelle der Interpretation einer latenten Variablen im Rasch-Modell nur die Interpretation einer Repräsentation der Verteilung der manifesten Item Responses heranziehen muss. Diese Überlegungen werden in Abschnitt 3.2 fortgeführt und es wird abgewogen, ob neben der typischerweise vorzufindenden Interpretation des Rasch-Modells als reflektives Modell auch eine Interpretation dieses Modells als formatives Messmodell möglich wäre. Diese Sichtweise wird insbesondere im Kontext von Kompetenzmessungen näher beleuchtet, in der die „Konstruktdefinition“ häufig nicht klar gefasst ist (bzw. gefasst werden kann).

In der Bildungsforschung werden im Large-Scale Assessment häufig komplexe Testdesigns (Multi-Matrix-Designs) mit mehreren Testheften eingesetzt, in denen Items an verschiedenen Positionen im Testheft auftreten. In Abschnitt 3.3 untersuchen wir anhand von Daten aus der Normierung der deutschen Bildungsstandards in der Grundschule, inwiefern Itemschwierigkeiten von der Position im Testheft abhängen (sog. Positionseffekte). In

einer weiterführenden Analyse wird festgestellt, ob die Kompetenz eines Schülers während des Tests relativ konstant bleibt oder individuelle Ermüdungseffekte (Persistenz) beobachtbar sind. Abschließend wird die Bedeutung der Zusammenstellung von Testheften mit Items aus verschiedenen Kompetenzbereichen diskutiert. Verschiedene Zusammenstellungen führen dabei zu sog. Kontexteffekten, die sich neben der mittleren Itemschwierigkeit in sog. Bookleteffekten niederschlagen. Zusammenfassend zeigen die Befunde in Abschnitt 3.3, dass in Multi-Matrix-Designs aufgrund von Positions-, Ermüdungs- und Bookleteffekten nicht von einer Itemschwierigkeit und einer Personenfähigkeit für alle Positionen im Testheft ausgegangen werden kann, sondern diese eher als „mittlere Parameter“ über alle im Testdesign repräsentierten Kontexte zu verstehen sind. Situationsspezifizität sollte daher explizit zugelassen bzw. in konkreten Testsituationen erwartet werden.

Eine häufige Verletzung des Rasch-Modells stellen lokale stochastische Abhängigkeiten dar, die in Abschnitt 3.4 diskutiert werden. Diese Abhängigkeiten können beispielsweise auftreten, wenn mehrere Items einen gemeinsamen Itemstimulus (ein Testlet) besitzen, beispielsweise, wenn mehrere Items zu einem Lesetext administriert werden. Schülerantworten für Items eines Testlets werden häufig stärkere Zusammenhänge als für Items in verschiedenen Testlets aufweisen. Dies führt zu zusätzlichen – im Rasch-Modell nicht modellierten – lokalen Abhängigkeiten. Diese lokalen Abhängigkeiten werden in einer Erweiterung des Rasch-Modells (dem sog. Testletmodell) als zusätzliche unkorrelierte Faktoren (Testletfaktoren) repräsentiert, deren Varianz (Testletvarianz) die Stärke dieser lokalen Abhängigkeiten charakterisiert. In Abschnitt 3.4 werden Testletvarianzen für einen Lesekompetenztest der deutschen Bildungsstandards in der Grundschule und für den deutschen Mathematiktest (DEMAT) bestimmt und deren Konsequenzen für die Kompetenzmessung diskutiert.

In Schulleistungsstudien finden Erhebungen meistens in Schulklassen statt, so dass Schüler in Klassen (und in Schulen) geschachtelt sind. Beim Überblick zu Item-Response-Modellen in Kapitel 1 fällt auf, dass die Modelle nur für Individuen und nicht für zwei Ebenen (Schüler und Schulklassen) formuliert sind. In Abschnitt 3.5 wird daher das Rasch-Modell mit sog. Multilevel-DIF-Modellen um klassenspezifische Itemschwierigkeiten erweitert. Für den Kompetenzbereich der Rechtschreibung in der Grundschule wird die Bedeutung des Multilevel DIF als Ausmaß der Variation von Itemschwierigkeiten zwischen Klassen untersucht. Es könnte sein, dass verschiedene Lerngelenheiten dazu führen, dass Itemschwierigkeiten zwischen Klassen variieren. Die Größe der Multilevel-DIF-Effekte im Bereich der Rechtschreibung wird abschließend mit Effekten im DEMAT verglichen.

In den Abschnitten 3.3, 3.4 und 3.5 werden drei mögliche Aspekte von Modellverletzungen im Rasch-Modell identifiziert und das Ausmaß dieser Verletzungen quantifiziert. Im Kapitel wird jedoch argumentiert, dass allein durch das statistische Aufdecken von „Effekten“ keine Notwendigkeit der Modellierung dieser Effekte in komplexeren IRT-Modellen nach sich ziehen muss.

2.2 Kapitel 4: Bedeutung der Itemauswahl und der Modellwahl in Längsschnittstudien

Kapitel 4 befasst sich mit der Erfassung von längsschnittlicher Veränderung am Beispiel der Lesekompetenz (dem ELFE-Test). Häufig werden Effektgrößen für die Beschreibung von Veränderungen verwendet.

Es wird in diesem Kapitel argumentiert, dass die statistische Unsicherheit dieser Effektgröße durch die Auswahl von Personen, die Auswahl von Items und die Wahl statistischer Modelle verursacht wird. In diesem Zusammenhang führt Kapitel 4 in das Konzept der Generalisierbarkeit von statistischen Parametern hinsichtlich der Facetten Personen, Items und Modelle ein.

Statistische Inferenz eines Parameters wird meistens im Hinblick auf die Generalisierung einer konkreten Stichprobe von Personen auf eine Population getroffen. Wie schon in Kapitel 3 angedeutet, muss in Schulleistungsstudien dabei auf die Mehrebenenstruktur der Schachtelung von Schülern in Klassen sowie Klassen in Schulen Berücksichtigung finden.

In Kapitel 4 argumentieren wir, dass die konkret in einem Test zur Messung längsschnittlicher Veränderung eingesetzten Items eine Quelle statistischer Unsicherheit darstellen, wenn sich die Generalisierung der Befunde der Studie auf eine größere Itempopulation beziehen sollen. Für jede Person kann eine längsschnittliche Veränderung verschieden ausgeprägt sein. Dies trifft jedoch auch auf Items zu: jedes Item, das für die Messung eines bestimmten Kompetenzbereiches eingesetzt wird, kann zu einer anderen längsschnittlichen Veränderung führen. In diesem Sinne sind berichtete Effektgrößen für Längsschnittdaten sowohl von der Auswahl von Personen als auch der Auswahl von Items abhängig. Für den ELFE-Test findet man, dass die Facette der Items gegenüber der Facette der Personen zu einem größeren „Standardfehler“ führt. Daher sollte für die statistische Inferenz in Längsschnittdaten auch die durch Itemauswahl bedingte Variabilität quantifiziert werden.

Für die statistische Modellierung längsschnittlicher Veränderung können verschiedene Item-Response-Modelle gewählt werden. Kapitel 4 zeigt im Rahmen des vorgeschlagenen Konzeptes der Generalisierbarkeit von Personen, Items und Modellen, dass die Wahl der Linkfunktion einen bedeutsamen Einfluss auf die Effektgröße längsschnittlicher Veränderung besitzt. Verschiedene Linkfunktionen strecken oder stauchen verschiedene Bereiche der Wahrscheinlichkeiten verschieden stark. Durch verschiedene Transformationen können damit auch itemspezifische Veränderungen verschieden gewichtet werden. Im Kapitel 4 wird der Standpunkt vertreten, dass Effektgrößen aus verschiedenen plausiblen Modellen berichtet und in einem statistischen Ansatz integriert werden sollten. Es wird dabei der Standpunkt vertreten, dass die „richtige“ Gewichtung von itemspezifischen Veränderungen nicht auf Basis eines Modellfits ermittelt werden sollte. Wenn keine theoretische Gründe für die Wahl eines IRT-Modells (und damit einer bestimmten Skalierungsmethode) vorliegen, so plädiere ich in Kapitel 2.2 dafür, die Variabilität von Effektgrößen aufgrund der Modellwahl als Teil der statistischen Inferenz mit Angabe von „Standardfehlern“ anzusehen.

2.3 Kapitel 5: Modellierung lokaler Abhängigkeiten

Dieses Kapitel widmet sich ausführlich der Fragestellung, ob und wie lokale Abhängigkeiten in IRT-Modellen Berücksichtigung finden sollten. In Abschnitt 3.4 wurden für einen Lesekompetenztest bereits Testleteffekte als Ausmaß der lokalen stochastischen Abhängigkeit berichtet und mögliche Konsequenzen diskutiert.

Häufig wird die adäquate Modellierung lokaler stochastischer Abhängigkeiten dadurch motiviert, dass korrekte Reliabilitätsschätzungen erhalten werden sollen. Daher wird in Kapitel 5 auf verschiedene Reliabilitätskonzepte eingegangen, die im Kontext lokaler Abhängigkeiten diskutiert werden.

Zunächst wird in Abschnitt 5.2 unterschieden, ob die im Test vorliegende Abhängigkeit als Teil der Traitvarianz oder der Fehlervarianz zu interpretieren ist. Es könnte bei C-Tests oder Lesekompetenztests der Fall sein, dass erst durch die Bearbeitung mehrerer Items zu einem Testlet alle Facetten des zu messenden Konstrukts aufgedeckt werden. Falls dies der Fall ist, so wäre das Vorliegen von Testlets konstruktinhärent und die Varianz als Traitvarianz (zumindest ein Teil davon) anzusehen. Dabei ist entscheidend, dass die statistische Modellierung der zusätzlichen Abhängigkeit von Items in Testlets dieselbe ist, je nach Interpretation würde man jedoch einerseits Abhängigkeit als Fehlervarianz (und damit einer geringeren Reliabilität), andererseits jedoch als wahre Varianz (und damit einer höheren Reliabilität) ansehen. Die Wahl der „richtigen Interpretation“ scheint aber nicht (ausschließlich) mit psychometrischen Methoden beantwortbar zu sein.

In Abschnitt 5.2 wird anschließend die klassische Reliabilität auf Basis einer Zerlegung einer beobachteten Kovarianzmatrix in eine Kovarianzmatrix der wahren Werte und eine Kovarianzmatrix der Fehler vorgestellt. Diese Matrixzerlegung ist ebenso tautologisch wie die Grundgleichung $X = T + E$ der klassischen Testtheorie. Für eine Bestimmung dieser Zerlegung sind bestimmte Annahmen notwendig. Häufig entsprechen diese Annahmen den Annahmen in Faktormodellen (d.h. Faktorenanalysen). Es wird gezeigt, dass sich die Bestimmung der klassischen Reliabilität unter der Annahme des Domain Samplings als eine spezielle Matrixzerlegung interpretieren und damit in das Konzept der klassischen Reliabilität überführen lässt. Dabei wird das Reliabilitätsmaß Cronbachs Alpha wird aus der Domain Sampling Perspektive motiviert, weshalb für die Berechnung von Alpha nicht die Annahme eines Faktormodells mit essentiell tau-äquivalenten Messungen notwendig erscheint.

Abschließend argumentiere ich in Abschnitt 5.2, dass auch fehlspezifizierte Faktormodelle zu korrekten Reliabilitätsschätzungen führen können. Dabei ist die Idee, dass sich die Modellfehler der Fehlspezifikation „ausmitteln“. Diese Annahme ist nicht statistisch testbar. Bezogen auf den Fall lokaler Abhängigkeiten heißt dies, dass auch die Anpassung eines Rasch-Modells unter Ignorierung der lokalen Abhängigkeiten legitimiert werden kann, wenn sich der theoretische Modellfehler (d.h. die nichtmodellierten Residualkovarianzen) zu Null addiert. Diese Annahme postuliert, dass man „im Mittel“ mit dem Rasch-Modell „die wesentliche Test-Dimension“ erfasst und sich positive und negative lokale Abhängigkeiten wechselseitig aufheben. Dieser Gedanke scheint für viele Kompetenzmessungen bzw. -modellierungen übertragbar.

In Abschnitt 5.3 diskutieren wir am Beispiel des Lesekompetenztests HAMLET verschiedene Arten der Modellierung lokaler Abhängigkeiten und es wird gezeigt, dass bedeut-

same Unterschiede in Reliabilitätsschätzungen auftreten. Außerdem wird herausgestellt, dass Modelle, die die Abhängigkeiten berücksichtigen, nicht mehr wie das Rasch-Modell nur auf dem ungewichteten Summenscore aller Items beruhen. Diese Eigenschaft besitzt sicher diagnostische Nachteile.

In Abschnitt 5.4 werden lokale Abhängigkeiten am Beispiel von Lesekompetenztests unter der Perspektive des Reliabilitäts-Validitäts-Dilemmas diskutiert. Dabei wird zunächst einer häufig anzutreffenden Behauptung widersprochen, dass die Reliabilität eine notwendige Voraussetzung der Validität darstellt. Dazu wird der Fall konstruiert, dass lokale Abhängigkeiten (und damit Testleteffekte) Teil der Validität sind, jedoch im Hinblick auf die Reliabilität Fehlervarianz darstellen. Dies führt dazu, dass unter Weglassen von Testlets bzw. Testleteffekten die Reliabilität steigt, die Validität jedoch sinkt.

Das Reliabilitäts-Validitäts-Dilemma wird abschließend im Kontext der optimalen Gewichtung von Testkomponenten bei Lesekompetenztests diskutiert. Dabei stehen einerseits Items zur Auswahl, die nicht in Testlets geschachtelt sind, und daher eine hohe Reliabilität, aber eine geringe Validität besitzen. Andererseits liegen in Testlets geschachtelte Items vor, die eher eine geringe Reliabilität, aber eine hohe Validität besitzen. In Abschnitt 5.4 wird eine Gewichtung beider Itemtypen in einem Test zur Maximierug der Validität vorgestellt.

Das Reliabilitäts-Validitäts-Dilemma verdeutlicht, dass die ausschließlich psychometrisch orientierten Reliabilitätsbetrachtungen oder Argumentationen für eine Wahl von IRT-Modellen mit Berücksichtigung lokaler Abhängigkeiten formal zwar die „richtige Reliabilität“ bestimmen kann. Diese Maßzahl drückt aber nicht hinreichend die „Präzision“ eines Tests aus. Dies wäre nur unter Rückgriff der Validität durchführbar, die aber in den meisten Anwendungen an Außenkriterien gekoppelt ist.

2.4 Kapitel 6: Item-Response-Modelle für fehlende Item Responses

In Kapitel 6 wird diskutiert, wie nicht bearbeitete Items bei der Skalierung von Leistungstests zu behandeln sind. Die in den übrigen Kapiteln diskutierten IRT-Modelle nehmen an, dass Item Responses nur unsystematisch gegeben der anderen beobachteten Item Responses ausfallen (d.h. missing at random; MAR).

In ausgewählter aktueller Literatur wird behauptet, dass fehlende Item Responses nicht als falsch, sondern tendenziell eher als ignorierbar zu bewerten seien. Es wird eine Präferenz geäußert, den Ausfallprozess modellbasiert zu behandeln. Ich setze mich in Kapitel 6 kritisch mit dieser Argumentation auseinander.

In Abschnitt 6.2 wird die probabilistische Modellierung von (beobachteten und nichtbeobachteten) Item Responses im Rahmen eines IRT-Modells (der sog. aleatorischen Unsicherheit) von der Unsicherheit des Rückschlusses eines fehlenden Item Responses auf eine „plausible“ Bewertung (der epistemischen Unsicherheit) unterschieden. Metaphorisch wird in ausgewählter Literatur mitunter argumentiert, dass man fehlende Item Responses nicht als falsch bewerten „darf“, da diese Datenbehandlung deterministisch sei, das IRT-Modell jedoch probabilistische Vorhersagen treffen würde. Ich zeige jedoch, dass diese Argumentation nicht stichhaltig ist, da aleatorische Unsicherheit von epistemischer Unsicherheit

zu unterscheiden ist. Außerdem wird eine Begründung auf Basis von Simulationsstudien für eine andere Behandlung der fehlenden Item Responses als die Bewertung als falsch in Frage gestellt, da diese Simulationsstudien bereits (trivialerweise) annehmen, dass die Bewertung als falsch nicht das datengenerierende Modell darstellt.

In Abschnitt 6.3 wird ein Überblick zu modellbasierten Verfahren der Behandlung fehlender Item Responses gegeben. Die Konsequenzen des Einsatzes bestimmter Modelle im Hinblick auf das Scoring der fehlenden Items werden herausgehoben. Es wird herausgestellt, dass die (meisten) modellbasierten Verfahren annehmen, dass der Ausfall auf einem Item nicht vom unbeobachteten Item selbst abhängt. Dies liefert für mich eine plausible Motivation, alternative IRT-Modelle ohne diese Annahme in Abschnitt 6.4 vorzuschlagen.

Ich schlage in Abschnitt 6.4 als erstes alternative IRT-Modell einen sog. Pseudo-Likelihood-Ansatz vor. Diese Modellierung ist als Sensitivitätsanalyse eines Parameters ρ zwischen den Extremen der Behandlung der fehlenden Item Responses als falsch ($\rho = 0$) und als ignorierbar ($\rho = 1$) anzusehen. Statistische Parameter können dann in Abhängigkeit des Sensitivitätsparameters ρ studiert werden, um den Einfluss verschiedener Annahmen an den Ausfallprozess auf interessierende Größen zu studieren.

Als ein zweites alternative IRT-Modell wird in Abschnitt 6.4 ein Modell vorgeschlagen, in dem neben der Fähigkeit und einer allen Items zugrundeliegenden Response-Tendenz auch der unbeobachtete Item Response einen Einfluss auf das Fehlen des Item Responses besitzt. Die Fälle der Behandlung der fehlenden Items als falsch und als ignorierbar stellen dabei Spezialfälle dar. Das vorgeschlagene IRT-Modell würde es sogar erlauben, den Ausfallmechanismus im Rahmen dieser Modellklasse aus den Daten zu identifizieren. Ob das aus Validitätsgründen jedoch in Anwendungen wünschenswert ist, darf bezweifelt werden.

Abschließend werden in Abschnitt 6.5 die Konsequenzen der Anwendung der verschiedenen IRT-Modelle zur Behandlung fehlender Item Responses für Ländervergleiche in PIRLS 2011 bei der Erfassung der Lesekompetenz illustriert. Es stellt sich heraus, dass die Art der Behandlung der fehlenden Item Reponses einen bedeutsamen Einfluss auf Unterschiede zwischen Ländermittelwerten besitzt.

2.5 Kapitel 7: Abschließendes Resümee

Kapitel 7 versucht, einige der in den Kapiteln 3 bis 6 aufgeworfenen methodischen Herausforderungen zu systematisieren.

Den Ausgangspunkt für Abschnitt 7.1 stellen die in Abschnitt 3.3 diskutierten IRT-Modelle zur Erfassung von Positions-, Ermüdungs- und Kontexteffekten dar. Dabei wird ein allgemeines IRT-Modell aufgestellt und gezeigt, wie sich die in Abschnitt 3.3 und aktuell in der Literatur vorgeschlagene Modelle darin einbetten lassen. Es wird festgehalten, dass Testleteffekte (vgl. auch Abschnitt 3.4 und Kapitel 5) und Ermüdungseffekte im Allgemeinen konfundiert sind. Die Konsequenzen von Positionseffekten für Längsschnittanalysen werden erörtert. Abschnitt 7.1 stellt heraus, dass es häufig unklar ist, ob Modellabweichungen von der lokalen stochastischen Unabhängigkeit (wie Testlet- oder Kontexteffekte) explizit modelliert werden müssen und daher zu verzerrten Parametern führen oder ob diese „nur“ Modellfehler darstellen, die zu keinen Verzerrungen in Para-

metern führen. Gegegebenenfalls müssen dann Standardfehler für Parameterschätzungen adjustiert werden.

In Abschnitt 7.2 wird die Bedeutung der Mehrebenenstruktur für IRT-Modelle diskutiert, da häufig Daten von Schülern vorliegen, die in Klassen oder Schulen genestet sind. Eine Möglichkeit ist dabei, die Mehrebenenstruktur als Störquelle zu betrachten und dann entsprechende Standardfehler für Itemparameter und Verteilungsparameter zu adjustieren. Alternativ können – wie in Abschnitt 3.5 vorgeschlagen – weitere hierarchische Verteilungen zur Modellierung der Mehrebenenstruktur eingeführt werden. Die in Abschnitt 3.5 eingeführten Multilevel-DIF-Modelle nehmen an, dass Itemschwierigkeiten zwischen Schulklassen (Multilevel DIF) variieren. In Abschnitt 7.2 die Möglichkeit erörtert, dass Multilevel DIF durch verschiedene Lerngelegenheiten verursacht sein könnten. Items mit starkem Multilevel DIF könnten dann besonders instruktionssensitive Items darstellen. Der querschnittliche Multilevel DIF wird in Zusammenhang mit vor kurzem in der Literatur vorgeschlagenen längsschnittlichen Multilevel DIF gebracht, der instruktional sensitive Items im Rahmen experimenteller Längsschnittstudien identifiziert. Abschließend wird der Fall beleuchtet, dass die Schülerebene eine Störquelle darstellt und das Messmodell auf ein Level-2-Konstrukt inferiert. Gerade bei Kompetenzmessungen für die Erstellung von Schulrückmeldungen könnte argumentiert werden, dass die interessierende Variable auf der Ebene von Klassen oder Schulen und nicht von Schülern zu verorten ist, da der Fokus der Rückmeldung hinsichtlich der Messung von Lernerträgen auf den höheren Ebenen liegt. Falls dies der Fall ist, sollten psychometrische Kriterien der Itemauswahl an Messmodellen für Level 2 orientiert sein.

In Abschnitt 7.3 werden Modellierungen und Interpretationen für den Fall einer unendlichen großen Itempopulation im Rahmen des Domain Samplings diskutiert. Dabei wird auf dem in Abschnitt 4.3 eingeführten Konzept der Generalisierbarkeit für die Facetten Personen, Items und statistische Modelle aufgebaut. Zunächst wird auf die Verbindung von Item-Response-Theorie und Generalisierbarkeitstheorie eingegangen. Es schließt sich eine Diskussion der Domain Sampling Perspektive für Reliabilitätsschätzungen und Faktormodelle an. Dabei grenzen wir Faktormodelle bzw. IRT-Modelle, die für eine unendlich große Itempopulation definiert sind, von statistichen Modellen ab, die explizit Modellfehler stochastisch modellieren. Die Möglichkeit des Zulassens von Modellfehlern im Rahmen Bayesianischer Faktormodelle wird herausgehoben. Abschließend diskutieren wir in Abschnitt 7.3 das Konzept der Invarianztestung unter der Perspektive des Domain Samplings. Ich argumentiere (wie auch schon in Kapitel 4), dass Invarianz (bzw. partielle Invarianz) keine Voraussetzung für die Untersuchung von Gruppenunterschieden darstellt, sondern nur eine von vielen anderen möglichen Identifikationsbedingungen ist. Es wird anhand der Fragestellung des Linkings aufeinanderfolgender PISA-Erhebungen herausgestellt, dass die Itemauswahl eine im Vergleich zur Stichprobenziehung von Personen bedeutsame Variabilitätsquelle sein kann.

Kapitel 3

Einige methodische Herausforderungen bei der Kalibrierung von Leistungstests

Dieses Kapitel geht auf einige methodische Herausforderungen bei der Kalibrierung von Leistungstests mit dem Rasch-Modell ein. In Abschnitt 3.2 werden verschiedene Interpretationen der latenten Variablen im Rasch-Modell eingeführt. Einige Modellalternativen zum Rasch-Modell werden in Abschnitt 3.1 diskutiert. Die restlichen Abschnitte dieses Kapitels befassen sich mit der Modellierung systematischer Modellabweichungen im Rasch-Modell. In Abschnitt 3.3 werden IRT-Modelle für Positionseffekte von Items bzw. Itemgruppen vorgestelt, mit denen abgeschätzt werden kann, ob die Schwierigkeit von Items (Itemgruppen) von ihrer Position im Testheft (Anfang vs. Mitte vs. Ende) moderiert wird und welche Implikation dies für Testungen hat. Es wird außerdem untersucht, ob die Effekte der Testheftposition intraindividuelle Varianz generiert. Der Abschnitt 3.4 behandelt Testleteffekte, die entstehen, wenn zu einem gemeinsamen Aufgabenstamm mehrere Items vorgelegt werden. Typischerweise wird dann die lokale stochastische Unabhängigkeit als Modellannahme des Rasch-Modells nicht mehr gegeben sein. Anhand zweier Beispiele wird illustriert, dass die Größe von Testleteffekten stark zwischen verschiedenen Lesetexten in einem Lesekompetenztest bzw. bestimmten Aufgabengruppen in einem Mathematiktest variiert. Abschließend werden in Abschnitt 3.5 differenzielle Itemcharakteristiken in Schulklassen (sog. Multilevel DIF) untersicht, die infolge differenzieller Lerngelegenheiten in unterschiedlichen Klassen entstehen können.

3.1 Alternativen zum Rasch-Modell

Werden in einem Test Items verschiedener Formate (Multiple Choice (MC), halboffene und offene Items) eingesetzt, so weisen vor allem MC-Items geringere Itemschwierigkeiten im Rasch-Modell auf. Damit gehen problematische Interpretationen von Items in verschiedenen Formaten für Kompetenzstufenmodelle einher, da ein Formatfaktor kein konstitutives Element von Kompetenzstufen darstellen sollte. Korrigiert man die Itemschwierigkeit um die Ratewahrscheinlichkeit, so fallen Schwierigkeitsschätzungen unabhängiger vom Itemformat aus. Kubinger und Draxler (2007) schlagen ein *Difficulty+Guessing Model* vor, das alle Itemtrennschärfen auf 1 setzt und itemspezifische Schwierigkeiten sowie Ratepa-

rameter annimmt und demzufolge einem restringierten 3PL-Modell entspricht. Sollten die Stichprobengrößen nicht notwendigen Anforderungen für eine Schätzung des Rateparameters genügen, so könnte auch der Rateparameter c_i auf $1/J$ fixiert werden, falls J der Anzahl der Antwortalternativen in MC-Items entspricht. Der Kritik, dass c_i in empirischen Anwendungen niemals genau $1/J$ betrage, kann entgegnet werden, dass im Gegensatz dazu im Rasch-Modell kein Rateverhalten modelliert wird (also $c_i = 0$ angenommen wird) und argumentiert werden muss, welcher der Rateparameter 0 oder $1/J$ plausibler ist. Alternativ schlägt de Gruijter (1986) vor, für alle Multiple-Choice-Items einen gemeinsamen Itemparameter zu schätzen, so dass hohe Anforderungen an die Stichprobengröße pro Item nicht notwendig sind.

Es muss an dieser Stelle betont werden, dass schlechte Schätzeigenschaften für den Rate- und Trennschärfeparameter bei moderaten Stichprobengrößen zeigen, dass sich in den Itemantworten zu wenig Informationen für eine Schätzung der Itemparameter ausschließlich mit Hilfe des betrachteten Items befinden. Dies heißt aber nicht zwangsläufig, dass das Rasch-Modell in diesen Situationen eingesetzt werden muss. Das Rasch-Modell setzt alle Itemtrennschärfen auf 1 und alle Rateparameter auf 0 und legt damit den Items starke Annahmen auf. Informative Priorverteilungen auf Itemparametern (Johnson & Albert, 1999; Kim & Bolt, 2007; Rupp, Dey & Zumbo, 2004) oder die Modellierung von Itemparametern als zufällige Effekte können extreme oder unzulässige Schätzungen vermeiden und stellen eine Regularisierung dar, ohne unplausible Annahmen im Rasch-Modell akzeptieren zu müssen (Ogasawara, 2002). Bayesianischen Verfahren (Gelman, Carlin, Stern & Rubin, 2004) fügen in Schätzmethoden Vorinformationen (Priorinformationen) über „plausible Werte“ von Parametern bereits ein. Im Fall der Item-Response-Modelle kann man beispielsweise annehmen, dass der Bereich plausibler Schätzungen bei MC-Items mit vier Antwortalternativen für einen Rateparameter normalverteilt mit einem Mittelwert 0.25 und einer Standardabweichung 0.1 ist, so dass bei moderaten Stichprobenumfängen hohe Rateparameter (etwa größer als 0.4) als unwahrscheinlich angesehen werden. Bayesianische IRT-Modelle sind mit der freien Software WinBUGS (Spiegelhalter, Thomas, Best & Lunn, 2003) relativ einfach umsetzbar und werden in der aktuellen Literatur stark diskutiert (Kim & Bolt, 2007; Patz & Junker, 1999a, 1999b; Sinharay, 2004). Diese Methoden werden für komplexere, stärker explorativ orientierte Modelle in diesem Kapitel eingesetzt.

Die ganzzahlige Schätzung von Itemladungen während der Itemkalibrierung ist mit dem vor allem in den nationalen Testungen in den Niederlanden verbreitete *One Parameter Logistic Model* (OPLM; Verhelst, Glas & Verstralen, 1995; Hoijtink & Vollema, 2003) möglich. Der Vorteil dieses Ansatzes besteht darin, die Itemparameterschätzung im Gegensatz zum 2-PL-Modell unabhängig von der Personenparameterschätzung mittels Conditional-Maximum-Likelihood-Verfahren wie im Rasch-Modell vornehmen zu können. Aus einer Bayesianischen Perspektive legt das OPLM diskrete Priorverteilungen auf die Trennschärfeparameter, so dass infolge der eingeschränkten ganzzahligen Schätzung extreme Schätzungen vermieden werden. Bei kleineren Stichproben kann den Daten in OPLM wie auch in Bayesianischen Ansätzen zur Verringerung der Variabilität der Parameterschätzungen ein höherer Grad an Regularisierung mittels stärkerer Priorinformationen auferlegt werden.

3.2 Zur Interpretation der latenten Variablen im Rasch-Modell

In diesem Abschnitt wird diskutiert, wie die latente Variable in einem Rasch-Modell zur Messung einer Kompetenz interpretiert werden kann. Dabei analysieren wir insbesondere die Kritik von Goldstein (1979) etwas näher und bringen diese mit der Unterscheidung formativer und reflektiver Messungen in Verbindung.

Goldstein (1979, S. 214) argumentiert, dass die Annahme der Eindimensionalität im Rasch-Modell für einen Itempool alle diskrepanten Items entfernt. Dabei ist nicht garantiert, dass die verbliebenen Items eine sinnvolle Skaleninterpretation erlauben. Die Eigenschaft doppelter Monotonizität im Rasch-Modell bedeutet, dass die Schwierigkeitsreihenfolge der Items für alle Schüler gleich ist, was gerade in Bereichen der Bildungsforschung aufgrund verschiedener Lernwege, Unterrichtsstile und Curricula mindestens für querschnittlich angelegte Messungen fragwürdig ist. Mit dem zusätzlichen Einsatz von Trennschärfeparametern im 2-PL-Modell würde diese Eigenschaft bekanntlich entfallen. Die alleinige Selektion von Items, die auf das Rasch-Modell passen („defining a test in order it fits a Rasch model"; Goldstein, 1979, S. 216), wird dabei in Frage gestellt. Willmott und Fowles (1974) bemerken dazu: „The criterion is that the items should fit the model, and not that the model should fit the items". Für praktische Kalibrierungsprozesse würde dies bedeuten, für die Itemmenge ein passendes IRT-Modell zu finden und nicht Items auszuwählen, so dass ein favorisiertes Modell „gut passt" (siehe dazu auch Hartig, 2008, S. 72). Üblicherweise stellt die Inspektion von Item-Fit-Statistiken (Infit und Outfit) mit vorgegebenen Toleranzbereichen Kriterien für die Itempassung des Rasch-Modells zur Verfügung. Die in Large-Scale-Assessments eingesetzten Faustregeln (etwa Infit kleiner als 1.1 oder ein Signifikanztest für den Infit-Parameter) scheinen eher konservativ vorzugehen, d.h. es werden tendenziell auch solche Items im Rasch-Modell beibehalten, die eigentlich eliminiert werden müssten. Goldstein, Bonnet und Rocher (2007) zeigen anhand der Daten von PISA 2000, dass die Annahme gleicher Itemladungen im Rasch-Modell und damit dessen Einsatz statistisch nicht haltbar ist. Die mitunter in der Literatur anzutreffende Praxis, dass nach Ausschluss einer gewissen Menge von Testitems aus dem gesamten Test (bezüglich eines Item-Fit-Kriteriums) das Rasch-Modell „gilt" und damit Eindimensionalität in den Daten gezeigt wird, ist kritisch zu hinterfragen. Eher scheint in diesem Fall eine Teilmenge von Items aus dem gesamten Instrument gefunden worden zu sein, auf die das Rasch-Modell unter vertretbaren Abweichungen hinsichtlich (lokaler) Item-Fit-Kriterien passt. Dieses rein statistische Vorgehen eignet sich aber sicher nicht zur Konstruktvalidierung (Baghaei, 2008).

Goldstein (1979, S. 219 ff.) verschiebt die Perspektive der Itemselektion für Anwendungen in der Bildungsforschung, indem er fordert „we can allow educational criteria properly to determine test content". Demzufolge würden also gewisse Testinhalte vordefiniert und in ein Instrument umgesetzt werden, wobei sich der Gesamttest als Gewichtung einzelner Bestandteile ergibt. Im Falle der Bildungsstandards kann beispielsweise ein Kompetenzbereich durch eine Menge von Standards mit relativen Gewichtungen (Bedeutung der Standards im Hinblick auf die gesamte Domäne) beschrieben werden und Itemmengen können entsprechend der Standards generiert werden. Die Skala für den gesamten Kom-

petenzbereich wird nach Goldstein damit aber sicher nicht mehr eindimensional sein. Die Standards mit ihren Operationalisierungen in Testitems *definieren* dann die Skala.

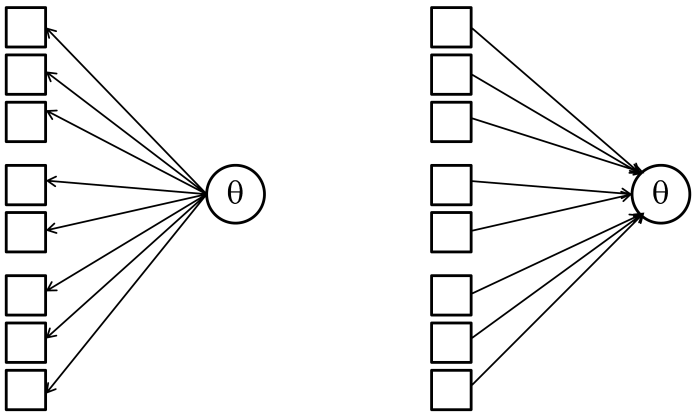

Abbildung 3.1: *Reflektives Messmodell (links) und formatives Messmodell (rechts)*

In Abbildung 3.1 werden die beiden verschiedenen kausalen Annahmen grafisch dargestellt. In der linken Grafik zeigen die Pfeile von der latenten Schülerfähigkeit θ zu den manifesten Itemantworten, d.h. eine Änderung in einer Schülerfähigkeit verursacht eine Änderung in den Itemantworten. Dieses Item-Response-Modell wird als *reflektiv* (Edwards & Bagozzi, 2000) bezeichnet, da die manifesten Items die latente Schülerfähigkeit reflektieren. Die rechte Grafik geht von einer *formativen* Messung aus: Die manifesten Items definieren die Schülerfähigkeit. Ursache und Wirkung sind also in beiden Messmodellen vertauscht. Gemäß unserer obigen Überlegungen konnten daher fur die Bildungsstandards formative Modelle angebracht sein, bei denen Testitems das Konstrukt definieren. Der Notation der Strukturgleichungsmodelle folgend entspricht dem Rasch-Modell die linke Grafik in Abbildung 3.1, bei der alle Pfeile mit einer Ladung 1 versehen sind. Allerdings ist die postulierte Beziehung zwischen Items und latenter Fähigkeit im Rasch-Modell nur korrelativer (assoziativer) und nicht kausaler Natur, so dass die Anwendung des Rasch-Modells sowohl unter Annahme reflektiver als auch formativer Messungen möglich ist (Stenner et al., 2008). Die geschätzte Schülerfähigkeit im Rasch-Modell stellt nur eine Transformation der Anzahl richtiger Items dar. Entsprechend konnten alle Items gleich gewichtet die Schülerfähigkeit in einem formativen Sinne bilden. Für Sijtsma (2006) sind latente Variablen in IRT-Modellen nur „Summaries“ der Daten (d.h. manifester Variablen): „latent variables [...] are summaries of the data and nothing more“ (Sijtsma, 2006, S. 452). Es findet demzufolge mit IRT-Modellen niemals eine Theorienprüfung, sondern immer nur eine Passung eines mathematischen Modells an die Daten statt. Einsichten in die den Schülerantworten zugrunde liegenden kognitive Prozesse oder die Gewinnung didaktischer Erkenntnisse werden mit (einfachen) psychometrischen Modellen nicht oder höchstens eingeschränkt möglich sein.

Goldstein (1979) nimmt bezüglich des Rasch-Modells eine Gegenposition zu Wright (1977) ein, der behauptet, dass in diesem Modell eine Gleichgewichtung der Items bedeutet, die Annahmen des Rasch-Modells (Eindimensionalität, gleiche Trennschärfen) akzeptieren zu müssen. Damit folgt Goldstein prinzipiell auch einer formativen Interpretation der latenten Variablen wie auch Stenner et al. (2008).

Die resümierende Forderung Goldsteins für die Itemkonstruktion in der empirischen Bildungsforschung scheint vielleicht auch noch 30 Jahre nach Erscheinen des Artikels Bedeutung zu besitzen (Goldstein, 1979, S. 220):

> I am arguing for a shift of emphasis away from a concern with the development of mathematical models aiming at neat technical solutions, and towards a development of quantitative assessment techniques which are firmly rooted in qualitative educational objectives.

Was bedeuten diese Überlegungen für Itemkonstruktionsprozesse, die Wahl von Messmodellen und die Interpretation latenter Variablen?

Für Kompetenzmessungen können nach Borsboom drei verschiedene Konzeptualisierungen vorgenommen werden (Borsboom, 2006, S. 434). Im Ansatz des *Universe-Sampling* der Generalisierbarkeitstheorie (Brennan, 2001a) werden sowohl Schüler als auch Items eines Tests als Stichprobe aus einer (hypothetisch unendlichen) Population von Schülern und Items verstanden. Konkrete Items in Bildungsstandardstests können dabei als Realisierungen möglicher Testitems verstanden werden, die von Aufgabenentwicklern (Didaktiker, Schulpraktiker, Psychometriker) mental individuell definiert oder mit einer durch einen gewissen Detailliertheitsgrad versehenen Beschreibung der Bildungsstandards in Verbindung gebracht wurden. Hartig (2008) betont, dass Items zur Kompetenzmessung, im Gegensatz zu Items zu theoretisch definierten psychologischen Konstrukten, in einem eng umgrenzten Bereich von Situationen definiert sind, aber gleichzeitig auch komplexes Verhalten abbilden sollen. Items zur Erfassung von operational definierten Kompetenzkonstrukten sind demzufolge im Vergleich zu psychologischen Konstrukten zu einem weitaus geringeren Ausmaß austauschbar (Hartig, 2008, S. 71). Borsboom und Mellenbergh (2007) zählen zum Universe-Sampling auch so genannte *Composite Attributes*, die im Beispiel von reinen Wissenstests (z. B. Faktenwissen) formativ einen Gesamttestwert bilden, aufgrund einer Stichprobenziehung von Items aus einer (großen) Itempopulation jedoch auf alle Items generalisieren. In diesem Vorgehen existiert damit nicht *der* Test, sondern Aussagen über Testergebnisse müssen im Hinblick auf eine Itempopulation getätigt werden. Insbesondere sind unter dieser Perspektive globale Modelltests für Messmodelle auf Itemebene wenig relevant, die von der Itemanzahl und damit auch von konkreten Items abhängen. Wird keine Sampling-Perspektive für die Items angenommen, so unterscheidet Borsboom *formative causal relations* (formative Modelle) von *reflective causal relations* (reflektive Modelle). Reflektive Modelle wie IRT-Modelle erlauben Modellgültigkeitstests. Formative Modelle hingegen definieren bereits den Zusammenhang zum Konstrukt und legen den Items im Messmodell demzufolge kein spezifisches Korrelationsmuster auf, sie können sogar unkorreliert sein (für weitere Entscheidungsregeln zwischen reflektiven und formativen Modellen siehe Jarvis, MacKenzie & Podsakoff, 2003, S. 203).

Fur Kompetenzmessungen wie in den Bildungsstandards schlagen wir ein so genanntes *Second Order Formative Model* vor, das in Abbildung 3.2 dargestellt ist. Dieses Modell wird auch als reflective first-order, formative second-order bezeichnet (Diamantopoulos, Riefler & Roth, 2008). Dabei wird angenommen, dass Items eine Menge latenter Variablen auf der ersten Ebene (first-order) reflektiv messen, d.h. diese Items können Operationalisierungen von latenten Variablen θ_1, θ_2 und θ_3 (etwa (Sub-)Standards oder Teilbereiche) darstellen. Bei hinreichender Spezifikation dieser Standards wird man im Allgemeinen es-

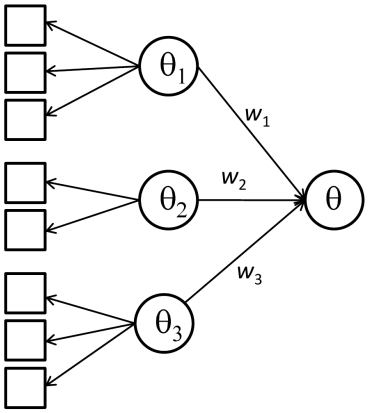

Abbildung 3.2: *Second Order Formative Model*

senzielle Eindimensionalität (siehe Böhme & Robitzsch, 2009) erwarten, so dass für Itemgruppen der Teilbereiche klassische Itemselektionskriterien Bedeutung besitzen könnten. Im Sinne der Überlegungen Goldsteins ist jedoch die gesamte Skala θ nicht zwingend eindimensional, sondern wird gemäß Gewichtungsanteilen w_1, w_2 und w_3 definiert, so dass auf der zweiten Ebene ein formatives Messmodell resultiert. Die Gewichtungen können im Test durch proportionale Itemanzahlen der Teilbereiche oder nachträgliche Verrechnungsvorschriften umgesetzt werden. Itemselektionsprozesse, die auf Eindimensionalität von θ abzielen, behalten eher homogene Items der Teilbereiche bei und decken demzufolge nicht die gesamte Breite des zu repräsentierenden Itempools zu Lasten der Validität ab. Berechnet man von den verbliebenen Items der Skala θ die Korrelation zwischen den Bereichen θ_1, θ_2 und θ_3, so werden im Allgemeinen Überschätzungen zu erwarten sein (siehe fur eine ähnliche Argumentation im Hinblick auf PISA-Items Goldstein, 2004).

Wenn Items in einer zu erfassenden Skala hinsichtlich der „richtigen Gewichtung“ (dies ist eine normative Setzung!) der Teilbereiche repräsentiert sind, dann ist die Anwendung eines Rasch-Modells im formativen Sinne eine sinnvolle Approximation des Second Order Formative Models als Messmodell. Itemfit-Statistiken würden dabei allerdings keine Bedeutung erlangen, ggf. würden also auch Items mit geringen Trennschärfen im Test beibehalten werden, wenn diese eine hinreichende Trennschärfe auf der Ebene der Variablen erster Ordnung (der Substandards) besitzen. Während die Eigenschaft der spezifischen Objektivität im Rasch-Modell hinsichtlich praktischer Relevanz kontrovers diskutiert wird (McDonald, 1999), sollten wichtige Vorteile der Anwendung dieser Modellklasse (ob in reflektivem oder formativem Sinn) hervorgehoben werden. Erstens erlauben Item-Response-Modelle (und insbesondere das Rasch-Modell) den robusten Umgang mit Multi-Matrix-Designs, so dass alle Schüler und alle Items auf eine Metrik gebracht werden können, obwohl im Allgemeinen nur wenige gemeinsame Items zwischen verschiedenen Testheften auftreten. Zweitens geht die Annahme gleicher Ladungen mit einer einfacheren Interpretation der Skala im Vergleich zu zweiparametrischen Modellen einher. Drittens liefert die Eigenschaft der doppelten Monotonizität im Rasch-Modell die gleiche Schwierigkeitsordnung aller Items unabhängig von der Schülerfähigkeit, so dass die Wahl der Normierung der Itemschwierigkeit auf eine bestimmte Lösungswahrscheinlichkeit (etwa 62.5% wie in PISA) die Itemordnung im Gegensatz zu 2-PL-Modellen nicht beeinflusst.

Schließlich stellen composite attributes gemäß Borsboom und Mellenbergh (2007) Beispiele für (second-order) formative Messungen dar, die im Hinblick auf Bildungsstandards eines Kompetenzbereiches eine Aggregation von Schülerfähigkeiten in einem interpretativ für breite Abnehmergruppen einfachen eindimensionalen Modell ermöglichen, das primär für kommunikative Prozesse ausgerichtet ist. Nicht zwingend muss dieses Modell (auch hinsichtlich der Dimensionalität) mit psychometrischen Modellen kohärent sein, auch wenn eine Interaktion der beiden Vorgehen für eine Spezifikation und Reflexion der Bildungsstandards auf empirischer Basis einige weitere Aspekte hervorbringen könnte.

3.3 Positions- und Bookleteffekte in Large-Scale-Assessments

In der Literatur zu den methodischen Herausforderungen von Large-Scale-Assessments (Wainer, 1993) besteht dahingehend Konsens, dass die Itemschwierigkeiten durch die Position im Testheft moderiert werden können und damit die Annahme der lokalen stochastischen Unabhängigkeit verletzt werden kann. Dies kann sich beispielsweise in *Testleteffekten* (vgl. Abschnitt 5 in diesem Kapitel) und *Bookleteffekten* (OECD, 2005, S. 195 ff.) niederschlagen, bei denen die Einbettung eines Items in ein bestimmtes Testheft oder eine Itemgruppe die Schwierigkeit beeinflusst.

Weiterhin treten solche Positionseffekte auf, je nachdem ob ein Item am Beginn oder am Ende eines Testhefts administriert wird. Typischerweise werden Items am Ende eines Testhefts schwieriger, häufig werden hierfür Ermüdungs- und oder Motivationseffekte auf Seiten der Testteilnehmer ins Feld geführt (Davis & Ferdous, 2005). Diese können umso stärker sein, je jünger die Testteilnehmer sind und je geringer der jeweilige Leistungsstand in der getesteten Domäne ist. Letzteres Argument ist in unserem Fall von hoher Relevanz, da in der Pilotierungsund Normierungsstudie neben Viert- auch Drittklässler immerhin zweimal 40 Minuten getestet wurden und weiterhin aufgrund der erheblichen Leistungsunterschiede innerhalb von Klassen davon auszugehen ist, dass schwächere Schüler besonders stark im Laufe der Testung ermüden.

Solchen unerwünschten Positionseffekten trägt man in Large-Scale-Assessments dadurch Rechnung, dass die Position von Items bzw. Itemgruppen variiert wird und sie zu Beginn eines Testheftes, in der Mitte und am Ende auftauchen. Die Position des Items stellt somit eine Kovariate dar, die den Itemantwortprozess beeinflussen kann. In folgender Gleichung wird dabei ein „positionsspezifisches Rasch-Modell“ im Sinne von *Explanatory Item Response Models* (De Boeck & Wilson, 2004) formuliert:

$$\text{logit}\left\{P(X_{pik}=1)\right\} = \theta_{pk} - b_{ik} \tag{3.1}$$

Formal wird dabei der Logit der Wahrscheinlichkeit einer richtigen Lösung eines Items i fur jeden Schüler p und fur jede Position k durch eine Fähigkeit (*Trait*, Kompetenz) θ_{pk} und eine positionsspezifische Itemschwierigkeit b_{ik} modelliert. Diese verschiedenen positionsabhängigen individuellen Fähigkeiten können als Methodenfaktoren zur Messung eines Merkmals interpretiert werden und die Position stellt eine Kovariate auf der Personenseite (*person side*; De Boeck & Wilson, 2004) dar. Zusätzlich kann auch die Itemschwierigkeit b_{ik} als über die Teftheftpositionen variierend angenommen werden, so dass Kovariaten auf

der Itemseite (*item side*; De Boeck & Wilson, 2004) existieren. Werden Items als Stichprobe aus einer Itempopulation angesehen, so verwendet De Boeck (2008) anstelle von Explanatory Item Response Models den Terminus *Random IRT Models*.

In Abschnitt 3.3.1 wenden wir uns Analysen zu, die nicht von positionsabhängigen Schulerfähigkeiten ausgehen und zusätzlich die Annahme treffen, dass Positionseffekte homogen auf alle Items wirken. Muss davon ausgegangen werden, dass der Positionseffekt auf Schuler differenziell wirkt, stellen wir in Abschnitt 3.3.2 eine Anwendung des obigen Modells dar, in dem die Traits θ_{pk} als lineare Funktion der Position modelliert werden.

Im Folgenden wollen wir solche Positionseffekte beispielhaft an eigenen Daten illustrieren. Dazu werden Analysen der Mathematikdaten von rund 9.000 Schülerinnen und Schülern aus der Normierungsstudie 2007 verwendet. Im Beitrag von Winkelmann und Böhme (2009) wird die Datenbasis ausführlicher vorgestellt. Das der Mathematik zugrunde liegende Kompetenzmodell wurde in Walther und Granzer (2009) ausführlicher vorgestellt.

Im Rahmen der Normierung der Mathematikitems zur Überprüfung der Bildungsstandards wurden Items derselben inhaltlichen Kompetenz immer in Blöcken gruppiert, deren Testzeit bei zehn Minuten lag. Als Folge konnten im Rahmen einer 80-minütigen Testung acht solcher Blöcke in einem Testheft auftauchen. Nach jeweils zwei Blöcken wurde die Testung kurz gestoppt und alle Schülerinnen und Schüler mussten danach einen neuen Block beginnen. Dadurch wurde gewährleistet, dass jeder Block bearbeitet werden konnte. Nach vier Blöcken (einer Schulstunde) erfolgte eine Pause von 15 Minuten.

Jeder Block wurde in unterschiedlichen Testheftversionen in seiner Position (Anfang, Mitte, Ende) variiert. Eine vollständige Permutation der Blöcke über alle acht Positionen hätte allerdings das Testdesign bzw. die möglichen Testheftvarianten zu komplex werden lassen. Jedoch war es möglich, die verschiedenen inhaltlichen Kompetenzen über alle acht Positionen zu variieren. Für die nachfolgend berichteten Befunde aus DIF-Analysen bedeutet dies, dass wir aggregiert über alle Items der jeweiligen inhaltlichen Kompetenz berichten, wie deren Schwierigkeitsparameter über die acht Positionen im Testheft variiert. Zusätzlich wurde nach Jahrgangsstufen (3 vs. 4) differenziert, um der Frage nachzugehen, ob die Positionseffekte bei jüngeren Schülerinnen und Schülern stärker als bei älteren ausfallen. Auf die inhaltliche Kompetenz „Daten, Häufigkeit und Wahrscheinlichkeit“ wurde wegen vergleichsweise geringer Itemzahlen verzichtet.

3.3.1 Positionseffekte auf der Itemseite

Für die weiteren Ausführungen nehmen wir an, dass die Position des Items im Testheft seine Schwierigkeit in gleichem Ausmaß für alle Personen beeinflusst. Um die Stärke solcher Effekte abschätzen zu können, werden oftmals Analysen zu differenziellen Itemfunktionen (*Differential Item Functioning*, DIF; vgl. Holland & Wainer, 1993) durchgeführt. Hierbei wird angenommen, dass der Schwierigkeitsparameter eines Items nicht (wie im Rasch-Modell vorausgesetzt) invariant ist, sondern in Abhängigkeit von der Position variiert. Für die Modellierung von Positionseffekten muss die Modellgleichung des Rasch-Modells um die Einführung von Parametern erweitert werden, welche die Positionen des Items kennzeichnen. Der Logit der Wahrscheinlichkeit für eine richtige Lösung des Items i an

Position k von Person p beträgt:

$$\text{logit}\left\{P(X_{pik}=1)\right\} = \theta_p - b_i - \nu_k \tag{3.2}$$

Dabei wird angenommen, dass die Summe der Positionseffekte ν_k über alle Positionen k gleich Null ist und die Fähigkeitsverteilung aller θ_p-Werte den Mittelwert 0 besitzt. Der Itemparameter b_i wird somit um den festen Wert ν_k in Abhängigkeit der Position k verschoben. Die Item-Response-Funktionen werden bildlich gesprochen in Abhängigkeit von der Position auf dem latenten Kontinuum verschoben. Dieses Modell kann in die allgemeine Modellklasse der LLTM-Modelle (Kubinger, 2008, 2009) eingebettet werden und eine Schätzung gemäß Conditional Maximum Likelihood ist möglich. Sind in der obigen Formel alle Positionseffekte ν_k (naherungsweise) gleich Null, so erhält man das gewöhnliche Rasch-Modell und die Modellierung der Position ist nicht notwendig.

In der vorliegenden Erhebung sind die Mathematikitems in acht Blöcken á zehn Minuten Bearbeitungszeit verteilt. Demzufolge werden im Rasch-Modell mit Positionseffekten acht Positionsparameter ν_1 bis ν_8 geschätzt, deren Summe 0 beträgt. Die Analysen wurden in WinBUGS (Spiegelhalter et al., 2003) unter Einsatz nichtinformativer Priorverteilungen mit 25000 Iterationen mit der Schätzmethode Markov Chain Monte Carlo (MCMC; Gelman et al., 2004) vorgenommen. Zwei der drei in diesem Kapitel hier vorgestellten Modelle können neben unserer Spezifikation in WinBUGS auch in ConQuest im Rahmen des *Multidimensional Random Coefficient Models* (MRCML, Adams, Wilson & Wang, 1997) geschätzt werden und es sind prinzipiell ähnliche Ergebnisse zu erwarten. Auch das Einfügen von multivariaten Itemkovariaten ist im Rahmen dieser Modellklasse möglich. Allerdings ist im MRCML keine Spezifikation von zufälligen Effekten auf Itemparametern (wie etwa in *Extended Generalized Linear Latent Mixed Models*; Segawa, Emery & Curry, 2008) möglich, wie dies im zweiten Modell dieses Abschnittes untersucht wird.

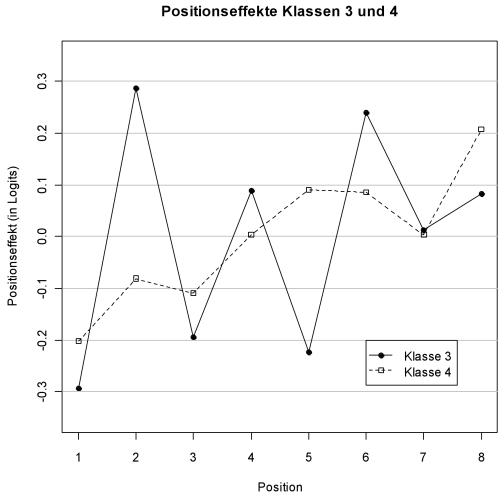

Abbildung 3.3: *Grafische Veranschaulichung der standardisierten Positionseffekte für Mathematikitems in den Klassenstufen 3 und 4*

Die Schätzungen der an der Standardabweichung der Schülerfähigkeiten standardisierten Positionseffekte sind in Abbildung 3.3 ersichtlich. Für die Klassenstufe 4 ergibt sich

ein näherungsweise linearer Verlauf der Positionseffekte von etwa -.2 Logits an der ersten Position (1. bis 10. Minute) bis zu .2 Logits an der 8. Position des Testes (71. bis 80. Minute). Beträgt die geschätzte Itemschwierigkeit eines Items zum Beispiel 1 Logit, so erwartet man nach dem Modell eine Itemschwierigkeit für dieses Item an der 1. Position von 1+(-0.2)=0.8 und für die 8. Position von 1+0.2=1.2. Daraus folgt, dass der Unterschied in den Itemschwierigkeiten einer standardisierten Mittelwertdifferenz von .4 Logits entspricht, etwa der Hälfte der Leistungsdifferenz zwischen der 3. und der 4. Klassenstufe im Fach Mathematik.

Es muss bei der Interpretation der Befunde beachtet werden, dass in der Testung die Schülerinnen und Schüler immer nach einer geraden Anzahl von Blöcken mit der Bearbeitung stoppen sollten und nach 20, 40 und 60 Minuten jeweils mit der Bearbeitung des darauf folgenden Blockes beginnen sollten. Dies erklärt in der 4. Jahrgangsstufe einen kleineren Positionseffekt für die dritte und siebte Position im Vergleich zur zweiten und sechsten Position.

Für die Klassenstufe 3 ergibt sich ein deutlich anderes Effektmuster. Auch in dieser Klassenstufe beträgt der standardisierte Unterschied etwa .4 (von -.3 der 1. Position bis zu etwa .1 der 8. Position), allerdings sind die Positionseffekte der Blöcke an geraden Positionen stets (deutlich) größer als an den davor liegenden ungeraden Positionen. Der Ermüdungseffekt innerhalb eines Paares von Blöcken tritt daher deutlich zu Tage. Der beobachtete Effekt ist nicht allein auf nicht erreichte Items (missing not reached) zurückführbar. Schülerinnen und Schüler der 3. Jahrgangsstufe bearbeiten Items zu Testbeginn oder zu Beginn eines neuen Blockes deutlich erfolgreicher. Die Ergebnisse weisen daher daraufhin, dass nicht von einem Itemparameter in einer Testbatterie gesprochen werden kann, sondern vielmehr von einem Itemparameter an einer bestimmten Position im Test. Zur Ermittlung eines „wahren Itemparameters“ ist deshalb der Einsatz an möglichst vielen Positionen im Testheft erwünscht, da sonst mit Verschätzungen der Itemschwierigkeit zu rechnen ist. In weiteren Studien ist zu prüfen, ob sich die beobachteten Befunde in diesem Ausmaß replizieren lassen.

Für Schüler der 4. Klassenstufe bei Mathematikitems der österreichischen Bildungsstandards zeigen Hohensinn et al. (2008) für einen Test von 40 Minuten eine Logitdifferenz vom ersten zum letzten Item von etwa .3 Logits, so dass sich die hier vorgestellten Befunde auch in anderen Assessments replizieren lassen. In Normierungsstudien auf der Grundlage der KMK Bildungsstandards Deutsch in der Sekundarstufe I ergaben sich für die Kompetenzbereiche Lesen sowie Sprachreflexion in 120 Minuten Testzeit Logitunterschiede der Positionseffekte, die den ermittelten Ergebnissen der 4. Jahrgangsstufe in Mathematik ähneln und damit die Befunde der Positionseffekte in PISA 2003 replizieren (siehe dazu Abschnitt 3.3.3 dieses Kapitels).

Itemspezifische Positionseffekte

Im vorgeschlagenen Modell wurde bislang angenommen, dass sich jede Itemschwierigkeit in Abhängigkeit der Position um den gleichen Wert ändert, die Items also homogen vom Positionseffekt betroffen sind. Schwierigkeiten weniger konzentrationssensitiver Items könnten allerdings weniger von der Position im Testheft abhängen. In folgender Modellgleichung wird daher positionsspezifische Heterogenität modelliert: die Itemschwierigkeit

für Item i an den Positionen $k = 1, \ldots, 8$ ergibt sich als itemspezifische lineare Funktion der Position:

$$\text{logit}\left\{P(X_{pik} = 1)\right\} = \theta_p - b_i - (k - 4.5)/7 \cdot \nu_i \tag{3.3}$$

Jedes Item i ist bezüglich seiner Schwierigkeit durch zwei Parameter b_i und ν_i charakterisiert. Dabei entspricht b_i der Itemschwierigkeit in der Testheftmitte, d.h. für diese Position ergibt sich mit $k = 4.5$ die Schwierigkeit $b_i - (k - 4.5)/7 \cdot \nu_i = b_i$. Der vom Item i abhängige Parameter ν_i drückt den itemspezifischen Positionseffekt aus. Die Kodierung der Positionen wurde dabei so gewählt, dass fur $k = 1$ in der Gleichung $(k - 4.5)/7 = -0.5$ folgt, fur k = 8 folgt als Kodierung der Position der Wert 0.5. Demzufolge entspricht der *Slope* (Anstieg, Positionseffekt) ν_i dem Anstieg der Itemschwierigkeit von der ersten zur letzten Position im Testheft. Zur Schätzung wird in WinBUGS eine zweidimensionale Normalverteilung der zufälligen Effekte b_i und ν_i spezifiziert. Eine Schätzung dieses Modells ist in ConQuest möglich, wenn b_i und ν_i als feste Parameter aufgefasst werden. Eine Bestimmung der latenten Korrelation zwischen Itemschwierigkeiten b_i und (linearem) Itempositionseffekt ν_i ist dann aber nicht mehr modellimplizit möglich.

Eine Schätzung des Modells fur die Schüler der 4. Klasse mit allen Mathematikitems der Normierung ergab einen Mittelwert der Verteilung der Positionseffekte ν_i von .37 und einer Standardabweichung von .59. Im Mittel beträgt der Unterschied der Itemschwierigkeiten von der ersten zur letzten Position in der 4. Klasse also .37 Logits. Die Korrelation zwischen Itemschwierigkeit und Itempositionseffekt beträgt -.05 und ist nicht signifikant. Es kann also anhand der Datenlage nicht belegt werden, dass Schwierigkeiten leichterer Items weniger stark von der Position im Testheft abhängen. Die substanzielle Standardabweichung der zufälligen Effekte ν_i zeigt jedoch, dass von heterogenen Positionseffekten auszugehen ist.

3.3.2 Positionseffekte auf der Personenseite

In Abschnitt 3.3.1 wurde angenommen, dass die Itemschwierigkeit um einen konstanten Positionseffekt verschoben wird. Items am Ende eines Tests werden demzufolge im Vergleich zum Beginn des Tests für alle Schüler in gleichem Ausmaß schwieriger. Hält man jedoch die Itemschwierigkeit konstant, bedeutet dies, dass Personen zu Beginn des Tests fähiger sind als zum Ende des Tests. Das Modell in Abschnitt 3.3.1 setzt allerdings voraus, dass der Fähigkeitsabfall in den Positionen des Testheftes für alle Schüler in gleichem Ausmaß erfolgt. Dies scheint in vielen Anwendungen jedoch unrealistisch: Schülerinnen und Schüler mit Konzentrationsschwierigkeiten werden deutlich eher ermüden, so dass die Homogenitätsannahme zu hinterfragen ist. In dem Modell

$$\text{logit}\left\{P(X_{pik} = 1)\right\} = \theta_{pk} - b_i - \nu_k \tag{3.4}$$

wird für jede der $k = 1, \ldots, 8$ Positionen eine Fähigkeitsdimension geschätzt. Zur Feststellung von Ermüdungseffekten ist dies jedoch kein einfach zu interpretierendes Modell, so dass wir für unsere Betrachtungen annehmen, dass die Fähigkeit eine lineare Funktion der Position im Testheft ist:

$$\text{logit}\left\{P(X_{pik} = 1)\right\} = \theta_p + (k - 4.5)/7 \cdot \varepsilon_p - b_i - \nu_k \tag{3.5}$$

Neben der interessierenden Schülerfähigkeit θ_p wird ein Ermüdungseffekt ε_k mit Mittelwert 0 eingeführt, der konstruktirrelevante Varianz enthält. Dieser Personenparameter ist nun der Slope der Fähigkeit, bei der die Testpositionen $k = 1, \ldots, 8$ linear auf den Bereich -.5 bis .5 rekodiert werden. Damit entspricht θ_p der individuellen Testleistung zur Mitte des Tests (Position 4.5; also zwischen 4. und 5. Block) und ε_p dem individuellen Unterschied zwischen der 1. und 8. Position nach Kontrolle des Positionseffektes über alle Schuler hinweg. Die Interpretation dieses Modells vereinfacht sich, wenn von Linearität der Positionseffekte ν_k von der Position k ausgegangen wird. Kürzt man die lineare Rekodierung der Testpositionen mit dem Term $a(k) = (k - 4.5)/7$ ab, so ergibt sich folgendes verallgemeinertes lineares Wachstumskurvenmodell (Embretson, 1991; Skrondal & Rabe-Hesketh, 2004):

$$\begin{aligned} \text{logit}\left\{P(X_{pik} = 1)\right\} &= \theta_p + a(k) \cdot \varepsilon_p - b_i - a(k) \cdot \nu \\ &= \theta_p + a(k) \cdot (\varepsilon_p - \nu) - b_i \end{aligned} \tag{3.6}$$

Der Slope der Fähigkeit (Fähigkeitsgradient) hat demnach über alle Schüler hinweg einen Mittelwert von $-\nu$, was dem mittleren Fähigkeitsunterschied zwischen der letzten und ersten Position im Testheft entspricht. Die individuellen Abweichungen vom mittleren Fähigkeitsgradienten $-\nu$ werden im Ermüdungseffekt ε_p erfasst. Wiederum wurde das Modell in WinBUGS spezifiziert.

Für die Bestimmung der Positionseffekte auf der Itemseite zeigt Abbildung 3.3, dass nur in der 4. Klasse die Annahme der Linearität des Positionseffektes plausibel ist. Es sollen daher exemplarisch nur Ergebnisse für drei Kompetenzbereiche aus der 4. Jahrgangsstufe in Tabelle 3.1 vorgestellt werden.

Tabelle 3.1: *Ergebnisse des latenten Wachstumskurvenmodells für Positionseffekte in drei inhaltlichen mathematischen Kompetenzbereichen der 4. Jahrgangsstufe*

	Zahlen und Operationen (I1)	Raum und Form (I2)	Größen und Messen (I4)
Position 1	−0.32	0.02	−0.21
Position 2	−0.04	−0.09	−0.07
Position 3	−0.16	−0.34	−0.05
Position 4	−0.03	0.05	0.08
Position 5	0.16	0.15	0.11
Position 6	0.14	−0.03	0.10
Position 7	−0.08	0.19	−0.23
Position 8	0.31	0.04	0.29
SD(Student)	1.08	1.09	1.15
SD(Slope)	1.05	1.65	0.85
Cor(Trait, Slope)	0.15	−0.30	0.20
Cor(Position, Positionseffekt)	0.77	0.47	0.52
Mean Slope	−0.39	−0.19	−0.23

In Tabelle 3.1 sind die festen Positionseffekte, die Standardabweichungen (SD) des Traits und des Ermüdungseffektes, der mittlere Ermüdungseffekt sowie die Korrelation

(Cor) zwischen Trait und Ermüdungseffekt abgetragen. Für den Kompetenzbereich „Zahlen und Operationen“ (I1) wird ein mittlerer Slope von -.39 geschätzt. Im Mittel beträgt demzufolge der über die lineare Regression geschätzte Leistungsfall vom Beginn zum Ende des Tests .39 Logits bei einer Standardabweichung von 1.05 Logits. Daraus berechnet sich ein Schwankungsintervall für den Ermüdungseffekt von $-.39 \mp 1.05 = [-1.44, 0.66]$, so dass in diesem Intervall etwa 68 Prozent der Ermüdungseffekte der Schüler zu erwarten sind. Im Vergleich zur Standardabweichung des Traits von 1.08 erscheint dieser Effekt sehr groß. Allerdings ist zu betonen, dass sich in der Methodenvarianz der Testposition auch spezifische Interaktionen von Schülerinnen und Schülern mit Aufgaben befinden. Sind also bestimmte Items oder Blöcke weniger stark von Ermüdungseffekten auf Seiten der Schülerinnen und Schüler betroffen, so schlägt sich dies bei variierenden Items pro Schüler in einer Interaktion von positionsspezifischer Fähigkeit und damit in einer Abweichung vom mittleren Ermüdungseffekt nieder. Eine von dieser Konfundierung bereinigte Interpretation des Ermüdungseffektes wäre nur dann gegeben, wenn alle Schüler die gleichen Items in der gleichen Reihenfolge vorgelegt bekommen hätten. Die leicht positive Korrelation zwischen Trait und Slope von .15 belegt, dass Schüler mit höherer Fähigkeit im Kompetenzbereich I1 („Zahlen und Operationen“) tendenziell größere Slopes und damit geringere Ermüdungseffekte aufweisen. Der mittlere Slope von -.39 wird also für bessere Schüler leicht in positiver Richtung erhöht. Schwächere Schüler besitzen demnach einen größeren negativen Fähigkeitsgradienten vom Beginn zum Ende des Testes. Ähnliche Befunde findet man auch für den Kompetenzbereich „Größen und Messen“ (I4). Davon abweichende Befunde werden jedoch für den Kompetenzbereich „Raum und Form“ (I2) beobachtet. Obwohl der mittlere Slope im Bereich I2 etwa so hoch wie im Bereich I4 ist, fällt bei großer Standardabweichung des Slopes von 1.65 die Korrelation von -.30 negativ aus. Bei höheren mittleren Schülerleistungen wird demnach der Ermüdungseffekt kleiner. Dies kann dahingehend interpretiert werden, dass in diesem Kompetenzbereich schwächere Schüler bei längeren Tests weniger benachteiligt sind als in den Bereichen I1 und I4.

Es ist zu betonen, dass diese Ergebnisse nur explorativen Charakter besitzen. In diesem Untersuchungsdesign ist es nicht möglich, die Vielzahl möglicher Methodeneffekte (Kontexteffekte, Testleteffekte, Positionseffekte) zugleich zu kontrollieren, was aber in einer komplexen Testsituation interpretativ angebracht erscheint (Mazzeo & von Davier, 2008). Vielmehr können diese Analysen leichte Evidenzen bieten, in Untersuchungen mit stärker auf diese Fragestellungen ausgerichteten Designs relevante Varianzquellen voneinander zu separieren.

3.3.3 Untersuchung von Bookleteffekten

Nachdem in den beiden letzten Unterabschnitten die Rolle der Position von Items im Hinblick auf Itemschwierigkeit und Schülerfähigkeiten untersucht wurde, soll nun der Einfluss der Vergabe eines Booklets (Testhefts) ermittelt werden. Aufgrund der Stichprobenziehung kann angenommen werden, dass die verschiedenen Booklets für den Mathematiktest zufällig den Schülern zugeordnet werden. In der Population dieser Schüler sollten demzufolge im Mittel keine Leistungsunterschiede in Mathematik beobachtbar sein. Sollten dennoch Leistungsunterschiede auftreten, so ist dies allein darauf zurückführbar, dass be-

stimmte Booklets aufgrund ihrer Anlage leichter oder schwieriger für die Schüler sind. Dieser Effekt hat nichts mit der mittleren Itemschwierigkeit im Booklet zu tun, nur die Komposition von Items und die Variation ihrer Kontexte verursacht ein differenzielles Funktionieren der Booklets. Für die Antwort eines Schüler p auf Item i in einem Booklet j nehmen wir folgendes Modell an OECD (2005, S. 198 ff.):

$$\text{logit}\left\{P(X_{pij}=1)\right\} = \theta_p - b_i - \nu_j \tag{3.7}$$

Der Parameter ν_j kennzeichnet dabei den Bookleteffekt fur Booklet j, der zusatzlich zu den Itemschwierigkeiten b_i die Wahrscheinlichkeit fur eine richtige Lösung des Items beeinflusst, wobei die Summe aller Effekte ν_j auf Null gesetzt wird. Fur den Mathematiktest zu den Bildungsstandards in der Normierung 2008 existieren folgende Booklet-Typen für die Klassenstufen 3 und 4:

1. fünf Booklets mit 40 Minuten Mathematik und anschließend 40 Minuten Deutsch
2. fünf Booklets mit 40 Minuten Deutsch und anschließend 40 Minuten Mathematik
3. zehn Booklets mit 80 Minuten Mathematik

Tabelle 3.2: *Bookleteffekte Mathematik in den Klassenstufen 3 und 4*

	Klasse 3		Klasse 4	
	M	SD	M	SD
Mathe/Deutsch	−0.10	0.23	−0.19	0.11
Deutsch/Mathe	−0.06	0.17	0.00	0.26
Mathe/Deutsch und Deutsch/Mathe	−0.08	0.19	−0.10	0.21
Mathe/Mathe	0.08	0.12	0.10	0.15

In Tabelle 3.2 sind Mittelwerte und Standardabweichungen der Bookleteffekte für die Booklet-Typen abgetragen. Dabei sind die Bookleteffekte in den Mathe/Deutsch-Booklets am geringsten ausgeprägt. Schüler in Klassenstufe 3, die zuerst 40 Minuten Mathematik und danach 40 Minuten Deutsch bearbeiten, erreichen im Mittel -.10 Logits, während Schüler die 80 Minuten Mathematik bearbeiten, .08 Logits erreichen. Dies entspricht etwa einem Unterschied von .2 Standardabweichungen. Dieser Effekt zeigt, dass Schülerinnen und Schüler in 40 Minuten Mathematiktestung bessere Leistungen erzielen als in 80 Minuten Mathematiktestung, da hier Ermüdungseffekte eine größere Rolle spielen. In Klassenstufe 4 ist der Unterschied mit fast .3 Logits noch etwas höher ausgeprägt. Interessant ist, dass die Mathematikleistungen für Schüler, die zuerst in Deutsch getestet wurden, geringer ausfallen als für die zuerst in Mathematik getesteten Schüler. Aber auch in diesem Fall sind die 40-Minuten-Testleistungen im Mittel besser als die 80-Minuten- Testleistungen. Die recht hohen Standardabweichungen der Bookleteffekte in Tabelle 3.2 indizieren jedoch, dass die Bookleteffekte spezifisch ausgeprägt sind. Die gemischten Booklets (Zeile „Mathe/Deutsch und Deutsch/Mathe“) variieren hinsichtlich der Bookleteffekte stärker als die reinen Mathematik-Booklets (SD=.19 zu SD=.12 in Klasse 3 bzw. SD=.21 im Vergleich zu SD=.15 in Klasse 4). Damit wird die Hypothese von Mazzeo und von Davier

(2008) bestätigt, dass Bookleteffekte bei Booklets mit verschiedenen Kompetenzen (*mixed designs*) stärker als bei Booklets mit nur einem Kompetenzbereich (*focused designs*) variieren. Bei den Deutsch/Mathematik-Booklets fällt auf, dass Booklets dann besonders schwer sind, wenn nur ein Kompetenzbereich in Deutsch (Schreiben oder Lesen) in den ersten 40 Minuten getestet wurde.

Die hier erhaltenen Ergebnisse lassen sich mit den in PISA berichteten Booklet- Effekten für die Sekundarstufe I in Verbindung bringen. In einem Mixed Design für die Lesekompetenz wird in PISA 2000 auf internationaler Ebene eine Standardabweichung der Bookleteffekte für die Lesekompetenz von .09 gefunden (Wu, Douglas & Monseur, 2002). Booklets, in denen Cluster mit Leseitems nur an zwei von vier Positionen auftreten, sind leichter als Booklets, in denen die Cluster an drei oder allen vier Positionen auftreten. In PISA 2003 mit Schwerpunkt auf der Erfassung mathematischer Kompetenz wurde ebenso ein Mixed Design eingesetzt. Die Standardabweichungen der Bookleteffekte für die erfassten Bereiche betragen: Mathematik SD=.20, Leseverstehen SD (OECD, 2005). Die Bereiche Lesen, Naturwissenschaften und Problemlösen werden dabei jeweils mit zwei Blöcken erfasst, die an allen vier Positionen im Testdesign auftreten und auf sieben Booklets verteilt sind. Nur in einem Booklet existierte dabei eine Verbindung zwischen den beiden Blöcken. Die für diese Bereiche ermittelten Bookleteffekte stellen aber Positionseffekte dar; die Unterschiede in den mittleren Leistungen auf den Blöcken von Position 1 zu Position 4 liegen bei $d = .48$ Standardabweichungen beim Lesen, $d = .57$ in Naturwissenschaften und $d = .40$ im Problemlösen, und sind durchaus als substanzielle Effekte anzusehen. Das Ignorieren dieser mittleren Unterschiede der Position von Itemblöcken führt im Allgemeinen zu einer Überschätzung der Varianz der Schülerfähigkeiten in den Kompetenzbereichen. Die Größenordnungen der Standardabweichungen der in den Bildungsstandards der Grundschule beobachteten Bookleteffekte liegen demzufolge in ähnlichen Größenordnungen, wie sie in PISA ermittelt wurden. Focused Designs oder Designs, in denen der Testablauf der Kompetenzbereiche in Mathematik standardisiert wäre (etwa die konstante Abfolge der Kompetenzbereiche in acht Blöcken: I1 – I2 – I3 – I4 – I5 – I1 – I2 – I4), sollte nach Mazzeo und von Davier (2008) Bookleteffekte verringern. Allerdings würde dabei eine Generalisierung von Itemparametern auf möglichst viele Positionen im Booklet nicht mehr möglich sein, was der Zielstellung in Normierungsstudien widerspräche.

3.3.4 Schlussfolgerungen

Die Itemnormierung für das Fach Mathematik umfasste eine Testzeit von zwei Schulstunden, in der Testblöcke in ihren Positionen variiert wurden. Die Befunde belegen, dass eine Vergleichbarkeit zukünftiger Erhebungen mit den vorliegenden Daten nur dann gegeben sein wird, wenn mit analogen Testzeiten und Testheftdesigns gearbeitet wird. Dies betrifft zunächst die zukünftigen Ländervergleiche im Primarbereich.

Bemerkenswert waren zweifelsohne die differenziellen Positionseffekte in beiden Klassenstufen. In der 4. Jahrgangsstufe nahm die Schwierigkeit mit steigender Position mehr oder weniger linear zu. Versteht man Positions- als Ermüdungseffekte, so nahm die Ermüdung über die Zeit hinweg kontinuierlich zu. Anders in der 3. Jahrgangsstufe, in der über die gesamte Testzeit abwechselnd deutliche Ermüdungsund Erholungseffekte beobachtbar

waren. Die Erholungseffekte ergaben sich immer dann, wenn die gesamte Gruppe gemeinsam mit einem neuen Testblock begann und damit zunächst Speed-Effekte ausgeschlossen werden konnten. Dies belegt, dass bei den jüngeren Schülerinnen und Schüler Speed- und Ermüdungseffekte auftraten. Schließlich zeigten die Befunde zu Positionseffekten auf die Personenparameter, dass tatsächlich schwächere Schülerinnen und Schüler anfälliger für derartige Ermüdungseffekte sind. Sollten doch Speed-Effekte vorliegen, so schlagen Glas und Pimentel (2008) ein zweidimensionales Item-Response-Modell vor, dessen erste Dimension die Richtigkeit der Itemantwort (Trait-Faktor) und die zweite Dimension eine Indikatorvariable für die Bearbeitung des Items durch den Schüler (Speed-Faktor) definiert[1]. Im Allgemeinen werden Trait-Faktor und Speed-Faktor korreliert sein.

Mit ihren Beschlüssen vom Juni 2006 hat die Kultusministerkonferenz beschlossen, zukünftige flächendeckende Vergleichsarbeiten in der Grundschule (VERA 3) auf der Basis der Bildungsstandards durchzuführen. Berücksichtigt werden sollen die Fächer Deutsch und Mathematik. Dazu sollen teilweise normierte Items und teilweise neu entwickelte Items verwendet werden. Die Testzeit pro Fach soll 60 Minuten nicht überschreiten. Konkret bedeutet dies, dass in Mathematik ein Itemblock nur die Positionen 1 bis 6 einnehmen kann. Damit ist zu erwarten, dass die Items leichter und die Personenfähigkeiten im Vergleich zur Normierungsstichprobe tendenziell höher ausfallen.[2] Ob solche Effekte allerdings in der Anwendung tatsächlich auftreten werden, bedarf der empirischen Überprüfung.

Aus diesen Analysen kann die Fragestellung der praktischen Implikation für Kompetenzstufenmodelle und der zugehörigen Fähigkeitsbeschreibungen für Schülerpopulationen resultieren. Aus Skalierungen gewonnene Itemparameter sind an eine konkrete Testsituation gebunden, d.h. beispielsweise Testlänge, Position im Test und der das Item umgebende Testkontext (andere Items, Instruktion, ...). Damit diese Parameter als Schätzung eines „Populationsmittelwertes“ dienen können, sollte der Einsatz eines Items in möglichst vielen Kontexten variiert werden (etwa Positionen, Kombination mit verschiedenen Blöcken). Wenn jedoch der Testablauf und damit auch die Reihenfolge der eingesetzten Items fixiert sind, scheint eine Bestimmung des Ausmaßes von Positionseffekten nicht von primärer Relevanz zu sein. Mazzeo und von Davier (2008) schlagen zur Verringerung von Kontexteffekten in Anlehnung an NAEP[3] vor, Focused Designs einzusetzen, bei denen in einem Testheft jeweils nur ein Kompetenzbereich getestet wird. Für die Ableitung von Kompetenzstufenmodellen, die empirisch an konkrete Testsituationen gebunden sind, stellt sich dann jedoch die Frage, ob Schüler verschiedene Instrumente zur Überprüfung mehrerer Kompetenzbereiche oder nur ein Instrument in der gesamten Testzeit bearbeiten sollen. Auch herrscht dann ein Bezug auf die Testzeit: „mittlere“ Itemparameter besitzen nur für 80 Minuten Testzeit Gültigkeit und ein Verhalten von Items in 40 oder 60 Minuten Testzeit kann ad hoc schwierig werden und sollte durch Brückenstudien erfasst werden.

Hinsichtlich der in diesem Abschnitt vorgenommenen Modellwahl ist anzumerken, dass die Annahme der homogenen Wirkung von Positionseffekten fur alle Items fragwürdig ist. In folgenden Studien kann beispielsweise die Hypothese geprüft werden, ob weniger kon-

[1] Siehe Kapitel 6 für eine vertiefte Diskussion zu IRT-Modellen mit fehlenden Item Responses.

[2] Abgesehen davon können auch die unterschiedlichen Bedingungen bei der Testdurchführung und -auswertung (durch das IEA Data Processing und Research Center (DPC) in der Normierung, durch Lehrkräfte bei VERA 3) zu nicht vergleichbaren Schätzungen führen.

[3] National Assessment of Educational Progress

zentrationsintensive Items (wie Items zum allgemeinen Kompetenzbereich „technisches Arbeiten und Routinentatigkeiten“) von geringeren Positionseffekten betroffen sind. Wenn in den bisher eingesetzten Modellen die Interaktion von Item i und Position k in der Schwierigkeit b_{ik} von der Zerlegung $b_i + \nu_k$ ausgegangen wird, können allgemeine (multivariate) Kovariaten des Items z_i außerdem in Bezug gesetzt werden, so dass die positionsspezifische Itemschwierigkeit b_{ik} als $b_i + \nu_k + \beta_k z_i$ modelliert werden könnte. Ergebnisse solcher Analysen könnten Informationen für „optimale Testkonstruktionen“ unter Verringerung von Positionssensitivitat geben, so dass beispielsweise weniger positionssensitive Items am Ende eines Tests administriert werden. Auch auf der Personenseite kann geprüft werden, wie Ermüdungseffekte mit anderen kognitiven Dispositionen oder Persönlichkeitsmerkmalen zusammenhängen.

Das in Abschnitt 3.3.3 vorgeschlagene Modell definiert die Schülerfähigkeit (formal) als Testleistung zum zeitlichen Mittelpunkt des Tests (also zur 40. Minute von 80 Minuten Testzeit). Man könnte diese Fähigkeit genauso auch als Baseline zu Beginn des Testes modellieren (Fähigkeit in den ersten Items des Tests) und damit den Abfall während des gesamten Tests feststellen. Dies erscheint jedoch aus Gründen unzureichender Reliabilität und der mangelnden Plausibilität (die Messung einer Fähigkeit ist ohne hinreichende Testzeit nicht möglich) kaum realistisch. Verallgemeinernd münden diese Überlegungen im Schluss, dass deskriptiv über die gesamte Testsituation von einer individuellen zunächst nichtparametrisierten Fähigkeitsfunktion (oder Fähigkeitstrajektorie; siehe für statistische Auswertungsmethoden dieser funktionalen Daten etwa Ramsay & Silverman, 2005) ausgegangen werden kann, die situations- und itemspezifische intraindividuelle Variabilität um eine konstante individuelle „Grundfähigkeit“ erfasst.

3.4 Testleteffekte: Zur Modellierung von Abhängigkeiten von Items mit einem gemeinsamen Stimulus

Für die Anwendung des Rasch-Modells wird (meistens) die lokale stochastische Unabhängigkeit als Modellvoraussetzung angesehen. Nach Kontrolle der individuellen Schülerfähigkeit sollen alle paarweisen Residualkorrelationen zwischen Items den Wert 0 besitzen. Gehören Items, wie im Bereich des Leseverstehens, zu einem gemeinsamen Stimulus (Lesetext), so spricht man von einem *Testlet*. In diesen Fällen ist im Allgemeinen von einer erhöhten positiven Abhängigkeit zwischen den Items eines Textes auszugehen. Zu verschiedenen Stimuli zugehörige Items sollten eine geringere Abhängigkeit aufweisen. In der Literatur können drei verschiedene Verfahren zum Umgang mit stochastischer Abhängigkeit unterschieden werden:

1. Trotz der Abhängigkeit werden IRT-Modelle eingesetzt, die auf Unabhängigkeit beruhen. Dies wird entweder damit begründet, dass die Missspezifikation keine Verzerrung relevanter Modellparameter verursacht oder damit, dass die Größe des Ausmaßes stochastischer Abhängigkeit ignoriert werden kann.

2. Alle Items mit einem gemeinsamen Stimulus werden zu einem Superitem zusammengefasst, das durch Summation aus allen Einzelitems entsteht. Dadurch wird die

Abhängigkeit zwischen den Items eines Stimulus aufgehoben (Swygert, McLeod & Thissen, 2001).

3. Es wird davon ausgegangen, dass Schüler mit Items eines gemeinsamen Stimulus interagieren, denn bestimmte Lesetexte „liegen“ manchen Schüler eher als anderen Schülern. In Testlet-Modellen (Bradlow, Wainer & Wang, 1999; Gibbons & Hedeker, 1992; Wainer, Bradlow & Wang, 2007) wird neben der primär zu erfassenden Kompetenz zusätzlich für jedes Testlet und jeden Schüler ein spezifischer Faktor eingeführt, der als Methodeneffekt der speziellen Methode des eingesetzten Stimulus interpretiert werden kann und die stochastische Abhängigkeit zwischen Items eines Stimulus quantifiziert. Li, Bolt und Fu (2006) vergleichen den Einsatz verschiedener Testlet-Modelle.

Für unsere Darstellungen setzen wir das *Rasch Testlet Model* (Wang & Wilson, 2005) als Spezialfall eines allgemeinen Testlet-Modells (Bradlow et al., 1999) ein. Das Rasch Testlet Model nimmt für alle Items identische Trennschärfen an:

$$\text{logit}\left\{P(X_{pit}=1)\right\} = \theta_p + \theta_{pt} - b_i \tag{3.8}$$

Die Wahrscheinlichkeit einer richtigen Lösung von Schüler p bei Item i im Testlet t setzt sich additiv aus der allgemeinen Fähigkeit θ_p und einem testletspezifischen Methodeneffekt θ_{pt} und der Itemschwierigkeit b_i zusammen. Es wird angenommen, dass alle Dimensionen (θ_p-Personeneffekte) unkorreliert sind. Im folgenden Datenbeispiel werden 16 Testlets modelliert, so dass insgesamt 17 unkorrelierte Dimensionen im Modell aufgenommen werden. Verschiedene Testlet-Varianzen sind dabei zugelassen. Je größer eine Testlet-Varianz ausfällt, desto größer ist die durch das Testlet induzierte lokale stochastische Abhängigkeit. Verschwindet in der obigen Formel der Term θ_{pt} für alle Testlets t, geht die Modellgleichung in die des Rasch-Modells mit lokaler stochastischer Unabhängigkeit über. Bei einer Testlet-Varianz von Null ist demzufolge neben der durch die allgemeine Fähigkeit θ_p verursachten Abhängigkeit keine (zusätzliche) Abhängigkeit zwischen den Items vorhanden.

Als Alternative zur Modellierung lokaler stochastischer Abhängigkeiten zur Einführung weiterer Personenparameter schlagen Hoskens und De Boeck (1995, 1997) die Einführung zusätzlicher Itemparameter für Itemgruppen vor, die lokale stochastische Abhängigkeit aufweisen. Auf ähnlichen Ansätzen beruhen Verallgemeinerungen von Multifacettenmodellen (Linacre, 1989), in denen Abhängigkeitsstrukturen von Schülerbeurteilungen bei mehreren Ratern durch weitere Itemparameter modelliert werden (Wilson & Hoskens, 2001). Im Ansatz so genannter marginaler Modelle werden neuerdings jedoch auch Copulas zur Erfassung von lokalen Abhängigkeiten in IRT-Modellen eingesetzt (Braeken, Tuerlinckx & De Boeck, 2007).

3.4.1 Testleteffekte für den Kompetenzbereich Lesen

In der Deutsch Nachnormierungsstudie (Primarbereich) im Frühjahr 2008 wurde der Schwerpunkt auf den Kompetenzbereich Lesen gelegt. Diese Zusatzstudie war nötig geworden, um Inkonsistenzen in den Itemparametern zwischen Pilotierungs- und Normierungsstudie zu erklären. Etwa 1600 Schülerinnen und Schülern der 3. und 4. Jahrgangsstufe

wurden zwölf Leseaufgaben verschiedener Textsorten vorgelegt. Einige Leseaufgaben treten in der Version der Pilotierung (P-Version) und der Normierung (N-Version) auf. Dabei unterscheiden sich die Stimuli der Aufgaben im Allgemeinen nicht, jedoch wurden einige Items in der Normierung weggelassen oder umformuliert.

Tabelle 3.3: *Testletvarianzen für 12 Leseverstehensaufgaben aus der Nachnormierungsstudie (2008)*

	Testletvarianz	
	P-Version	N-Version
zwischen Schülern	1.11	
Text 1		0.57
Text 2	0.34	0.22
Text 3	0.27	
Text 4	0.64	0.40
Text 5	0.02	0.15
Text 6	0.62	
Text 7		0.92
Text 8		3.60
Text 9	0.23	
Text 10	2.58	
Text 11		0.32
Text 12	0.67	0.98

Die entsprechenden Analysen auf der Basis des Rasch Testlet Models wurden mit WinBUGS durchgeführt. In Tabelle 5 sind die Varianzen der allgemeinen Fähigkeit des Leseverstehens und die Testlet-Varianzen aufgeführt. Die Größe der Testlet-Varianz ist jeweils an der Varianz der Schülerfähigkeit von 1.11 zu relativieren. Beispielsweise werden für Text 2 relativ geringe Testlet-Varianzen von .34 und .22 in beiden Versionen beobachtet. Bei diesem Text ist das Antwortverhalten der Schüler wenig spezifisch auf die Interaktion von Schüler und Text zurückführbar. Die qualitative Analyse der Aufgaben mit sehr großen Testlet-Varianzen (Texte 8 und 10) zeigt, dass die eingesetzten Items bezüglich der Aufgabenstellung stark von den übrigen Items im Test abweichen (diskontinuierliche Textstimuli; Items, die auch zum Kompetenzbereich „Sprachgebrauch untersuchen“ gezählt werden können). Insgesamt unterscheiden sich die Testlet-Varianzen bei den Aufgaben in zwei Versionen zwar statistisch signifikant, sie sind aber hinsichtlich praktischer Relevanz in ähnlichen Größenordnungen einzuschätzen.

3.4.2 Diskussion

Implikationen für die Skalenkonstruktion

In diesem Kapitel wurde demonstriert, wie sich Testleteffekte in IRT-Modellen einfach analysieren lassen. Die teilweise recht großen Testleteffekte sind dadurch begründet, dass die in der Analyse berücksichtigten Lesetexte aus der Nachnormierungsstudie auch Items

enthalten, die man dem Kompetenzbereich „Sprachgebrauch untersuchen“ zuordnen kann. Insgesamt können hinter der Varianz von Testlets verschiedene Quellen verborgen sein: Abweichungen bezüglich der Aufgabenstellung von der Hauptdimension Leseverstehen, differenzielle Schülerfähigkeiten in verschiedenen Textsorten, differenzielle Schülermotivation hinsichtlich bestimmter Themen oder differenzielle Schülerperformanz aufgrund verschiedener Positionen einer Aufgabe im Testheft.

Die Größe der Testlet-Varianz kann ohne Frage als Kriterium bei der Aufgabenselektion eingesetzt werden. Bei (normativ feststellbaren) ähnlichen Iteminhalten ist einem Lesetext Vorrang zu geben, der geringere Abhängigkeiten zwischen Items und damit eine kleinere Testlet-Varianz besitzt. Ist aber Heterogenität beispielsweise durch verschiedene Textsorten des Leseverstehens hinsichtlich der Konstruktdefinition vorgegeben und ist demzufolge nach Berücksichtigung dieser Textsorten von jeweils eindimensionalen Subskalen auszugehen, dann wird die Gesamtskala formativ gebildet (Edwards & Bagozzi, 2000; vergleiche dazu auch die formative Interpretation des Rasch-Modells). Hohe Testlet-Varianzen können dann hinsichtlich aller im Test auftretenden Texte unterrepräsentierten Textsorten entsprechen, so dass eine spezifische Textsorte weniger im Gesamtfaktor und dafür umso mehr im Testlet-Faktor sichtbar wird. Aus psychometrischer Sicht sind dann hohe Testlet-Varianzen oder geringe Itemtrennschärfen (die in IRT-Skalierungen in ConQuest häufig großen Infit- oder Outfit-Werten entsprechen) eine direkte Folge der *Skalendefinition*. Diese Größen haben dann eher deskriptiv informativen Charakter und müssen nicht zwingend zu Item- oder Aufgabenausschluss führen. Damit geht aber die Abschwächung der Forderung von Eindimensionalität oder maximaler Skalenreliabilität einher, die im Fall formativer Messungen kein notwendiges Kriterium darstellt.

Implikationen aus statistischer Perspektive

Neben dem hier eingesetzten Rasch Testlet Model werden in der Literatur Testlet- Modelle vorgeschlagen, die sowohl Trennschärfe als auch Rateparameter beinhalten (Bradlow et al., 1999). In folgender Modellgleichung wird neben der Itemschwierigkeit eine itemspezifische Trennschärfe angenommen:

$$\text{logit}\left\{P(X_{pit}=1)\right\} = a_i \cdot (\theta_p + \theta_{pt} - b_i) \tag{3.9}$$

Dieses Modell lasst sich unter der Annahme, dass die Varianz des primaren Faktors θ_p gleich 1 ist, identifizieren. DeMars (2006) untersuchte im Rahmen einer Simulationsstudie, welche Auswirkungen der Einsatz eines fehlspezifizierten Modells (2-PL-Modell und Modell mit einem Superitem fur jedes Testlet) besitzt. Die Reliabilität wird bei Annahme des 2-PL-Modells anstelle des Testlet-Modells über-, bei Einsatz eines Modells mit Superitems unterschätzt. Damit ist der übliche Einsatz des Rasch-Modells bei existierenden Testleteffekten hinsichtlich der Reliabilität als zu liberal einzuschätzen. Sowohl im 2-PL-Modell als auch im Testlet-Modell werden die Itemschwierigkeiten b_i erwartungstreu geschätzt, im Testlet-Modell ist die Parameterschätzung im Vergleich zum 2-PL-Modell jedoch (etwas) effizienter. Die Itemtrennscharfen a_i werden allerdings im 2-PL-Modell unterschätzt. Dies entspricht einer kleineren extrahierten Varianz des Generalfaktors und zeigt sich deutlich in Anwendungen mit sehr stark ausgepragten Abhängigkeitsstrukturen zwischen Items (etwa bei C-Tests). Sind jedoch vor allem Itemschwierigkeiten von Interesse, so ist

nicht von einer substanziellen Beeinflussung der Ergebnisse bei der Vernachlässigung von Testleteffekten auszugehen.

Brandt (2008) hat die Annahme der bedingten Unkorreliertheit von Testlet-Faktoren infrage gestellt. Beispielsweise könnten bei der Erfassung von Leseverstehen Texte mit ähnlichen Themen (etwa Sachtexte zu verschiedenen Haustieren) zu größeren Abhängigkeiten zwischen Items dieses Bereiches als zwischen Items verschiedener Bereiche (etwa Sachtexten und literarischen Texten) führen. Demzufolge sei die Unkorreliertheit der Testlet-Faktoren abzuschwächen. Sein vorgeschlagenes *Rasch Subdimension Model* (Brandt, 2008) stellt allerdings nur die Umparametrisierung eines mehrdimensionalen Rasch-Modells dar, das genauso viele Dimensionen wie Testlets besitzt. Damit ist dieses Modell unseres Erachtens schon bei einer moderaten Anzahl von Testlets nicht mehr handhabbar. Wir halten es jedoch für sinnvoll, so genannte Second Order Testlet Models für eine mögliche Verletzung bedingter stochastischer Unabhängigkeit einzusetzen. Neben der allgemeinen Lesekompetenz werden einige Testlets mit ähnlichen Eigenschaften (etwa Inhalten oder Textsorten) als unkorrelierte Testletgruppen aufgefasst und zusätzlich (residuale) Testlets modelliert. Damit stellt dieser Ansatz ein Mehrebenenmodell (Gelman & Hill, 2007) dar, bei dem Items in Testlets geschachtelt und Testlets wiederum in Testlet-Gruppen geschachtelt sind.

Neben der Testlet-Varianz als Maß für die lokale stochastische Unabhängigkeit innerhalb eines Testlets kann jedoch auch der Mittelwert der Residualkorrelationen über alle Items eines Testlets als Maßzahl verwendet werden (Yen, 1984 setzt die so genannte Q_3-Statistik ein). Je positiver die Residualkorrelationen eines Testlets ausfallen, desto höher ist das Ausmaß der lokalen stochastischen Abhängigkeit. Bei Vorliegen eines Multi-Matrix-Designs muss aber die Q_3-Statistik und die Höhe der Testlet-Varianz nicht zwangsläufig hoch korrelieren, wenn nicht jeder Schüler für die Gesamtdimension Lesen einen repräsentativen Itempool bearbeitet und nicht alle Blöcke von Items in allen Blockkombinationen untereinander verbunden sind. Je mehr Testlets jedoch von einem Schüler bearbeitet werden, desto höher werden Q_3 und Testlet-Varianzen korrelieren.

Ebenso kann mit der nichtparametrischen Methode DETECT nach Vorgabe von aus Testlets bestehenden Itemgruppierungen die durch Testlets induzierte Abweichung von Eindimensionalität im Sinne essenzieller Dimensionalität (Stout, 1987) quantifiziert werden (Jang & Roussos, 2007; Winkelmann & Robitzsch, 2009).

3.5 Multilevel DIF: Modellierung klassenspezifischer Itemschwierigkeiten

Für jeden Schulleistungstest stellt sich die Frage von Testfairness und curricularer Validität. Gibt es Items, die bestimmte Klassen oder Bundesländer spezifisch bevorteilen oder benachteiligen? Bestimmte Iteminhalte oder Itemformate könnten im Unterricht bestimmter Klassen häufiger geübt worden sein (*Opportunity-to-Learn*; Muthén, Kao & Burstein, 1991), so dass diese Items für die betreffenden Klassen tendenziell einfacher als für andere Klassen sind. Damit funktionieren diese Items in starkem Ausmaß differenziell über Schulklassen hinweg. Während in traditionellen Analysen zu „klassischem“ *Differential Item Functioning* (DIF; Holland & Wainer, 1993) differentielles Itemfunktionieren für ei-

ne geringe Anzahl fester Gruppen (z.B. Geschlecht oder soziale Herkunft) untersucht wird, unterscheidet sich die Frage von Multilevel DIF für Klassen dadurch, dass die Klassen der Stichprobe aus einer größeren Population von Klassen gezogen werden und demzufolge nur die Varianz des DIF über die Klassen hinweg von Interesse ist. Variiert die Itemschwierigkeit dem zufolge auch nach Kontrolle von Schülerfähigkeiten sehr stark über die Klassen hinweg, so handelt es sich um ein klassenkontextsensitives Item, das aus Gesichtspunkten der Testkonstruktion vom Einsatz in Vergleichsstudien vielleicht eher auszuschließen ist.

Das im Folgenden eingesetzte *Multilevel DIF Rasch Model* (Chaimongkol, 2005; Chaimongkol, Huffer & Kamata, 2007; Kamata & Cheong, 2007) ist ein spezielles hierarchisches IRT-Modell (Fox & Glas, 2001; Goldstein et al., 2007; Johnson, Sinharay & Bradlow, 2007). Der hierarchische Modellcharakter manifestiert sich dabei nicht ausschließlich dadurch, dass Schüler in Klassen genestet sind. Vielmehr bedeutet hierarchisch, dass Teilmengen von Modellparametern als aus einer Verteilung stammend modelliert werden (Gelman & Hill, 2007). Die Wahrscheinlichkeit für eine korrekte Itemlösung von Schüler p in Klasse s auf Item i wird modelliert gemäß

$$\text{logit}\left\{P(X_{psi}=1)\right\} = \theta_s + \theta_{ps} - b_{is}\ ,\ E(b_{is}) = b_i\ , Var(b_{is}) = \sigma_i^2 \tag{3.10}$$

Dabei ist θ_s die mittlere Kompetenz der Klasse s, θ_{ps} die Abweichung der Kompetenz des Schülers p in Klasse s vom Klassenmittelwert θ_s und b_{is} die Itemschwierigkeit des Items i in Klasse s. Fur die Schätzung des Modells nimmt man an, dass alle Mittelwerte der θ-Werte 0 betragen und die Verteilung der Itemschwierigkeiten b_{is} normalverteilt mit Mittelwert b_i und einer gemeinsamen Varianz σ_i^2 ist. Der Parameter b_i entspricht dann näherungsweise der geschätzten Itemschwierigkeit in einem Rasch-Modell unter Nichtberücksichtigung der Mehrebenenstruktur.

Je größer die Varianz $\sigma_i^2 = Var(b_{is})$ ausfällt, desto stärker ist für dieses Item ein klassenspezifisches differenzielles Funktionieren zu beobachten. Zur Feststellung der Signifikanz dieser Varianz (Nullhypothese H0: Varianz ist gleich Null) wird mit den gewonnenen Modellparametern (Itemschwierigkeiten, Intraklassenkorrelation) aus einer Umsetzung in WinBUGS (Spiegelhalter et al., 2003) ein neuer Datensatz unter der der Annahme simuliert, dass kein Multilevel DIF existiert, also ein Mehrebenen-Rasch-Modell gilt. Für diese generierten Daten wird das Multilevel DIF Modell geschätzt und kritische Werte werden für Varianzen des Multilevel DIF (etwa 90%- Quantile aus der Posteriorverteilung der Varianzparameter) abgeleitet. Diese Werte sind demnach für die Multilevel DIF Parameter zu erwarten, wenn in der Population kein Multilevel DIF vorliegt. Zur Beurteilung der praktischen Relevanz der Multilevel DIF Standardabweichung als Effektgröße scheint jedoch die Angabe eines *Bayesian Credibility Intervals* (Gelman et al., 2004) sinnvoll. Das Bayesian Credibility Interval kann man aus empirischen Quantilen der Posteriorverteilung der Multilevel DIF Standardabweichung gewinnen, welche wiederum einfach aus dem WinBUGS-Output verfügbar sind.

Anstelle der Reprasentation der klassenspezifischen Itemschwierigkeiten b_{is} durch eine Normalverteilung mit einem Mittelwert und einer Varianz nimmt Vermunt (2008) eine diskrete Mischverteilung fur diese Schwierigkeiten an. Unser eingesetztes Modell besitzt nicht das Problem der Definition der Anzahl der Dimensionen und erscheint uns inhaltlich plausibler.

Wir illustrieren die Untersuchung von Multilevel DIF anhand zweier Beispiele aus dem Bereich der Rechtschreibung und der Mathematik.

3.5.1 Multilevel DIF bei einem Rechtschreibtest der Bildungsstandards

In der Deutsch-Normierungsstudie des Primarbereichs (2007) wurde als Rechtschreibtest ein Lückendiktat eingesetzt (Böhme & Bremerich-Vos, 2009). Gerade bei dieser Testform ist die Untersuchung von Multilevel DIF von Relevanz, da davon ausgegangen werden kann, dass bestimmte Wörter in Klassen mehr oder weniger stark geübt werden könnten. Im Folgenden stellen wir Ergebnisse für ein aus 20 Items bestehendes Lückendiktat dar, das in 4. Klassen eingesetzt wird. Mit Hilfe der aus dem Multilevel DIF Modell gewonnenen Varianzen ermittelt man eine Intraklassenkorrelation von .21, so dass 21% der beobachteten Varianz der Rechtschreibleistungen auf die Klassenzugehörigkeit der Schülerinnen und Schüler zurückführbar ist. Tabelle 3.4 gibt für 20 Items einer ausgewählten Aufgabe einen Überblick über die Itemschwierigkeit und die Standardabweichung der Itemschwierigkeiten (über Klassen) als Maßzahl für Multilevel DIF. Die Standardabweichung des Traits der Rechtsschreibung bei dieser Aufgabe beträgt 1.32. Alle Standardabweichungen (SD) des Multilevel DIFs größer als .44 sind statistisch signifikant größer als Null auf dem Niveau .10, eine SD des Multilevel DIF größer als .51 ist statistisch signifikant auf dem Niveau .05.

Das leichteste Item „Vogelfutter“ (Item 16) mit einer mittleren Itemschwierigkeit von -1.74 besitzt mit einer Standardabweichung der Itemschwierigkeiten von 0.72 einen signifikanten Multilevel DIF. Damit liegen etwa 68% aller klassenspezifischen Itemschwierigkeiten im Schwankungsintervall $-1.74 \mp 0.72 = [-2.46, -1.02]$. Bei diesem Item variiert also die Schwierigkeit über Klassen hinweg substanziell. Sehr große Klasseneffekte hinsichtlich der Schwierigkeit besitzen außerdem die Wörter „kaputt“ (Item 6), „vorspielen“ (Item 9), „eigentlich“ (Item 10) und „flitzen“ (Item 11). Die hohe Multilevel DIF Standardabweichung von 3.78 des Wortes „flitzen“ bei einer überraschend hohen Schwierigkeit weist jedoch nach genauerer Analyse der Schülerlösungen auf ein testadministratives Problem hin. In vielen Klassen verstanden die Schüler beim Diktieren fälschlicherweise „flitzten“ (oder die Testleiter diktierten falsch), so dass die große Variabilität der Itemschwierigkeit zwischen Klassen erklärbar ist.

Es bleibt zu untersuchen, ob wichtige Prädiktoren auf Itemebene (z.B. das Item befindet sich nicht im Grundwortschatz der untersuchten Klasse, Lehrereinschätzungen zu Lerngelegenheiten bestimmter Wörter, etc.) die Höhe der Multilevel DIF Varianzen erklären können.

3.5.2 Multilevel DIF beim Mathematiktest DEMAT

Als weitere Anwendung des Multilevel DIF untersuchen wir die Abhängigkeit der Itemschwierigkeit vom Klassenkontext anhand des Mathematiktests DEMAT4 (Gölitz, Roick & Hasselhorn, 2006) für die 4. Klasse. Dieser Test umfasst die in den Bildungsstandards formulierten inhaltlichen Kompetenzen außer dem Bereich „Daten, Häufigkeit und Wahrscheinlichkeit“. In der Mathematik Normierungsstudie ergab sich für diesen Testteil bei

Tabelle 3.4: *Multilevel DIF für ein Lückendiktat in Klasse 4 aus der Normierungsstudie (2007)*

			Multilevel DIF		
		Mittlere Item-	SD Item-		
		schwierigkeit	schwierigkeit	σ_i	σ_i
Wort	Item	$b_i = E(b_{is})$	$\sigma_i = SD(b_{is})$	10%-Quantil	90%-Quantil
Lehrerin	1	−1.51	0.37	0.07	0.68
versteckt	2	−1.05	0.48	0.09	0.83
Verkehr	3	0.11	0.15	0.02	0.32
gesperrt	4	1.76	0.31	0.06	0.56
Staubsauger	5	−1.04	0.34	0.07	0.63
kaputt	6	0.81	0.73	0.51	0.97
Geburtstag	7	−0.24	0.49	0.19	0.79
Geschenke	8	−1.28	0.57	0.24	0.89
vorspielen	9	−0.59	0.99	0.70	1.33
eigentlich	10	−0.58	1.10	0.82	1.42
flitzen	11	2.05	3.78	2.84	4.86
Schlittschuhläufer	12	1.48	0.52	0.26	0.77
empfehlen	13	1.30	0.39	0.13	0.65
Kartoffelsuppe	14	−0.23	0.37	0.10	0.62
verstaubten	15	0.83	0.33	0.08	0.56
Vogelfutter	16	−1.74	0.72	0.38	1.06
unübertrefflich	17	0.97	0.28	0.08	0.52
verabschiedete	18	0.88	0.19	0.05	0.38
Schlüsselloch	19	−0.26	0.41	0.14	0.67
Wohnzimmer	20	−1.58	0.22	0.04	0.46

Anmerkung: Wörter von statistisch signifikanten Multilevel DIF Effekten sind fett dargestellt. Aus den 10%- und 90%-Quantilen gewinnt man ein Bayesian Credibility Interval zum 80%-Niveau.

Anwendung des Multilevel DIF Modells eine Intraklassenkorrelation von .25. Insgesamt 40 Items sind in neun Bereiche klassifiziert worden, die aus Abbildung 3.4 hervorgehen. In jeder Zeile dieser Abbildung befindet sich die Multilevel DIF Standardabweichung mit zugehörigem 80% Bayesian Credibility Interval. Dabei fällt deutlich auf, dass die Items zum Bereich „schriftliche Division“ den größten Multilevel DIF aufweisen. Möglicherweise führen verschiedene Übungsintensitäten oder unterrichtliche Schwerpunktsetzungen in diesem Bereich zum differenziellen Funktionieren dieser Items. Ebenso sind die Items zu schriftlichen Multiplikationen, Spiegelungen und Größenvergleiche in besonderem Ausmaß klassenspezifisch vom DIF betroffen. Diese Bereiche können als „leicht trainierbar“ angesehen werden und wären damit von der Intensität des Opportunity-to-Learn betroffen. Einen gering ausgeprägten Multilevel DIF beobachtet man im Bereich „Lagebeziehungen“ und Sachrechnen, die entsprechend unspezifisch hinsichtlich der Items im Unterricht repräsentiert sind. Möglicherweise sind die Items zur schriftlichen Addition und schriftlichen Multiplikation in Klasse 4 weniger von Multilevel DIF betroffen, weil diese beiden Bereich im Unterricht der 3. Klassenstufe zu verorten sind und die Übungsintensität sich nicht in Klasse 4 niederschlägt.

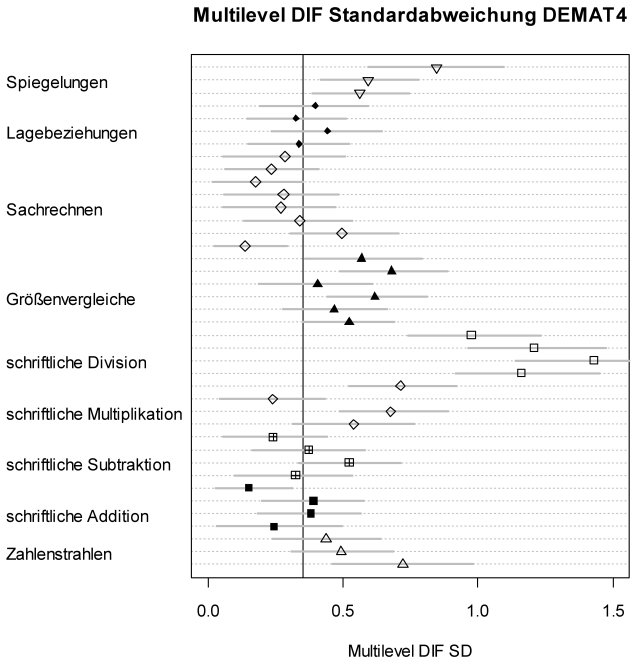

Abbildung 3.4: *Multilevel DIF Standardabweichungen für Items des DEMAT4. Gleiche Symbole in der Abbildung bedeuten den gleichen Aufgabenbereich. Die vertikale Linie bei 0.35 symbolisiert den kritischen Wert der Standardabweichung für einen signifikanten DIF-Effekt auf dem 10%-Konfidenzniveau. Zusätzlich sind die 80% Bayesian Credibility Intervals abgetragen.*

Die Befunde zum Multilevel DIF werden prinzipiell auch für den DEMAT3 in der dritten Jahrgangsstufe repliziert. Die höchsten DIF-Varianzen der Items besitzen die Aufgabenbereiche „schriftliche Subtraktion“, „schriftliche Multiplikation“ und „Spiegelungen“, so dass auch bei diesem Test Items mit hohem Anteil an Routinetätigkeiten in den Schwierigkeiten am stärksten zwischen Klassen variieren.

3.5.3 Simultane Betrachtung von Testleteffekten und Multilevel DIF

Die Untersuchung des Multilevel DIF für DEMAT4 legt jedoch nahe, verschiedene Varianzquellen näher zu differenzieren. Da alle Items des Bereiches „schriftliche Division“ von starkem Multilevel DIF betroffen sind, wird untersucht, ob dieser Klasseneffekt nicht homogen für eine Klasse für alle Items des Aufgabenbereiches wirkt, d. h. für bestimmte Klassen sind damit tendenziell alle Items im Vergleich zum Durchschnitt leichter oder schwieriger. Die einzelnen Aufgabenbereiche im DEMAT4 interpretieren wir dafür im Folgenden als Testlets, die neben der allgemeinen Mathematikkompetenz als unkorreliert angenommen werden.[4] Für einen Schüler p in Klasse s auf einem Item i einer Aufgabe

[4]Diese Annahme ist natürlich nur näherungsweise gültig, da die Aufgabenbereiche der Arithmetik (schriftliche Addition, Subtraktion, Multiplikation und Division) untereinander höher korrelieren als mit Bereichen der Geometrie (etwa Spiegelungen). Die Erfassung der vollen mehrdimensionalen Struktur unter Berücksichtigung der Mehrebenenperspektive würde das Modell allerdings an dieser Stelle unnötig ver-

(Testlet) t modellieren wir zunächst

$$\text{logit}\left\{P(X_{psit}=1)\right\} = \tilde{\theta}_{ps} + \tilde{\theta}_{pst} - b_{is} \tag{3.11}$$

Diese Gleichung entspricht dem Testlet-Modell in Abschnitt 3.4 mit zusätzlichen Itemschwierigkeiten b_{is}, die zwischen den Klassen s variieren. Sowohl die allgemeine Fähigkeit $\tilde{\theta}_{ps}$ als auch die bereichsspezifischen Faktoren $\tilde{\theta}_{pst}$ werden in einen Klasseneffekt θ_s bzw. θ_{st} und einen Individualeffekt θ_{ps} bzw. θ_{pst} zerlegt:

$$\text{logit}\left\{P(X_{psit}=1)\right\} = \theta_s + \theta_{ps} + \theta_{st} + \theta_{pst} - b_{is} \tag{3.12}$$

Dieses Modell integriert damit die Zweiebenen-Perspektive, den Testleteffekt sowie Multilevel DIF und fällt in die Modellklasse der Extended Generalized Linear Latent Mixed Models (Segawa et al., 2008). Gleichung (3.12) kann als so genanntes *IRT Variance Decomposition Model* (van den Berg, Glas & Boomsma, 2007) aufgefasst werden, das eine Varianzkomponentenzerlegung in verschiedene Variationsquellen auf der Logitmetrik vornimmt. Die Generalisierbarkeitstheorie (*G Theory*; Brennan, 2001a) hingegen setzt bei dichotomen Rohdaten auf Itemebene an und führt die Varianzzerlegung auf der Metrik der Rohwerte durch, so dass eine „Konfundierung der Varianzquellen“ mit Itemschwierigkeiten erfolgt. Briggs und Wilson (2007) nutzen Item-Response-Modelle, um auf der Metrik der Rohwerte eine Varianzkomponentenzerlegung wie in der G-Theory einzusetzen. Die Modellgleichung (3.12) fällt in die Modellklasse der kreuzklassifizierten Mehrebenenmodelle (Van den Noortgate et al., 2003). Die Verbindung von Testleteffekten und Multilevel DIF Fragestellungen in Anwendung auf die Vergleiche internationaler Survey-Fragebogenskalen modellierten in einem ähnlichen Modell unter zusätzlicher Berücksichtigung von Itemtrennschärfen de Jong, Steenkamp und Fox (2007).

Auch das Modell (3.12) ist in WinBUGS (Spiegelhalter et al., 2003) verhältnismäßig einfach spezifizierbar. Die Höhe der verschiedenen Varianzquellen aus obiger Gleichung liefert aus der Sichtweise der Generalisierbarkeitstheorie Einblicke in klassenspezifisches Itemfunktionieren und klassenspezifische Stärken in bestimmten Domänen der mathematischen Kompetenz.

In Tabelle 3.5 sind in den ersten beiden Spalten die Standardabweichungen des Generalfaktors Mathematik θ_s bzw. θ_{ps} in DEMAT4 und der Testlet-Faktoren θ_{st} bzw. θ_{pst} abgetragen. Die gesamte Standardabweichung (3. Spalte) ergibt sich als Wurzel der Summe der Varianzen auf Schüler- und Klassenebene. Das Verhältnis von Klassenvarianz zur Gesamtvarianz ist als Intraklassenkorrelation (ICC, 4. Spalte) gekennzeichnet und gibt an, in welchen Bereichen besonders große Unterschiede zwischen Klassen zu verorten sind. Beim Generalfaktor sind 26% der Varianz der Mathematikleistungen auf Klassenunterschiede zurückführbar, während die auf Testlet-Ebene residual definierten Intraklassenkorrelationen nur bei der schriftlichen Division mit ICC = .42 einen sehr hohen Wert annehmen. Nur noch die Bereiche „schriftliche Subtraktion“, „Größenvergleiche“ und „Spiegelungen“ weisen auf recht hohe klassenspezifische Varianzanteile in den Testlets hin. Bis auf den etwas geringer ausgeprägten Effekt im Testlet „schriftliche Multiplikation“ kann damit durch diese Befunde gezeigt werden, dass die in Abschnitt 3.5.2 ermittelten Effekte zum großen

komplizieren.

Tabelle 3.5: *Ergebnisse des Modells mit Testleteffekten und Multilevel DIF für DEMAT4*

	Standardabweichungen				
Faktor	Klasse	Schüler	Gesamt	ICC	Testlet SNR
Generalfaktor	0.99	1.69	1.96	.26	1.00
Zahlenstrahlen	0.28	2.69	2.70	.01	0.53
schriftliche Addition	0.54	3.03	3.08	.03	0.41
schriftliche Subtraktion	1.30	3.84	4.05	.10	0.23
schriftliche Multiplikation	0.69	2.66	2.75	.06	0.51
schriftliche Division	2.73	3.19	4.20	.42	0.22
Größenvergleiche	0.68	1.89	2.01	.11	0.95
Sachrechnen	0.25	2.15	2.16	.01	0.83
Lagebeziehungen	0.48	3.87	3.90	.02	0.25
Spiegelungen	1.29	3.86	4.07	.10	0.23

Teil auf ein differenzielles klassenspezifisches Funktionieren im gesamten Bereich zurückgeführt werden können. In der letzten Spalte der Tabelle wird durch den Testlet *Signal To Noise Ratio* (*Testlet SNR*) das (Gesamt-) Varianzverhältnis von Generalfaktor und Testlet-Faktor angegeben. Niedrige Werte (wie Testlet SNR = .22 im Bereich „schriftliche Division“) bedeuten, dass die Items in diesem Aufgabenbereich stark voneinander abhängig sind und dimensional vom Generalfaktor stärker abweichen. Nicht zwingend muss jedoch – wie die Ergebnisse in diesem Beispiel zeigen – ein korrelativer Zusammenhang zwischen Testlet SNR und ICC bestehen.

Die Standardabweichungen der in der Modellgleichung verbliebenen Multilevel DIF Effekte σ_i korrelieren nur moderat zu .45 mit den Multilevel DIF Effekten aus dem Modell ohne Testleteffekte in Abschnitt 3.5.2. Dies stützt die Interpretation, dass Multilevel DIF kein singulär itemspezifisches Phanomenen darstellt.

3.5.4 Diskussion

Es muss betont werden, dass sich Multilevel DIF Effekte hinreichend stabil in Klassen identifizieren lassen, wenn eine Mindestanzahl von Schülern innerhalb einer Klasse Items zugleich bearbeitet haben. Dies widerspricht der Zielstellung des Large- Scale-Assessments, eine große Itemmenge über die verschiedenen Schüler in den Klassen zu streuen, um Clustereffekte in Itemparameterschätzungen zu minimieren. Sollte in Assessments die Untersuchung von Multilevel DIF von Bedeutung sein, so müsste in Teildesigns die Rotation von Testheften innerhalb von Klassen gering ausgeprägt sein und die Möglichkeit der Erfassung des Opportunity-to-Learn auf Lehrer- und Schülerseite gegeben sein. Nicht nur Klassen, sondern auch Bundesländer können aufgrund verschiedener Schwerpunktsetzungen in Lehrplänen oder der Unterrichtskultur Gegenstand für Multilevel DIF zur Prüfung interessanter Hypothesen darstellen. Items mit einem starken Multilevel DIF sind möglicherweise besonders gut dafür geeignet, die Effizienz von Unterrichtsprozessen zu evaluieren, da diese Items besonders sensitiv für den Klassenkontext und ggf. weniger sensitiv für außerschulische Lerngelegenheiten sind.

Bei einer Betrachtung von Multilevel DIF über 16 Bundesländer hinweg würde man das Problem umgehen, für eine Untersuchung differenziellen Itemfunktionierens nicht die Länder in wenige Gruppen (etwa südliche und nördliche oder östliche und westliche Bundesländer) klassifizieren zu müssen. Auf der Ebene von Items kann dann untersucht werden, welche Itemkovariaten Multilevel DIF Varianz vorhersagen können. Beispielsweise könnten spezifische Kompetenzbereiche oder Itemformate einen gewissen Teil der DIF Varianz zwischen Bundesländern erklären. Solche Analysestrategien werden in kommenden Arbeiten vorgestellt.

3.6 Abschließende Bemerkungen

Abschließend sollen einige resümierende Bemerkungen in zwei Unterpunkte angebracht werden: die Mehrebenenperspektive in IRT-Modellen und die Problematik der (psychometrisch orientierten) Modellwahl.

3.6.1 Multilevel IRT-Modelle

Die in diesem Kapitel vorgestellten IRT-Modelle illustrieren den komplexen Mehrebenencharakter, insbesondere wenn sowohl Schüler als auch Items als Stichproben aus Populationen aufgefasst werden (Brennan, 2006; De Boeck, 2008). Die Schüler sind in Klassen und diese wiederum in Bundesländern geschachtelt. Items selbst können in Testlets, inhaltlichen Domänen oder aber in Itemfamilien strukturgleicher (oder paralleler) Items geschachtelt sein (Johnson et al., 2007). Von dieser Stichprobenziehungsperspektive kann aber der Einsatz von varianzkomponentenanalytischen Überlegungen differieren. Durchaus können auch kreuzklassifizierte Beziehungen wie die Interaktion von Schüler und Testlet (Abschnitt 3.4 zu Testleteffekten) oder von Klassen mit Items (Abschnitt 3.5 zu Multilevel DIF) relevante Fragestellungen betreffen und Ausgangspunkt explorativer Analysen sein. Das Zusammenwirken verschiedenster Situationen des Assessments wird durch die Erfassung der verschiedenen Variationsquellen (insbesondere von Interaktion) aus der Perspektive der Generalisierbarkeitstheorie genauer beleuchtet. Brennan (2006) bemerkt, dass in Assessments hinterfragt werden müsse, was genau eine Replikation eines Tests ausmacht. Er kritisiert die in den meisten IRT-Modellen vorherrschende Annahme von festen Items, wonach sich eine Bestimmung der Standardfehler und der Reliabilität auf eine exakte Replikation des Tests mit denselben Items beziehen würde. In diesem Kapitel unterstrichen wir jedoch die Auffassung, dass die Idee des Universe-Sampling Stichprobenfehlers bei der Itemselektion explizit modelliert werden sollte.

3.6.2 Zur Modellwahl

Die Wahl adäquater Analysemodelle stellt für Assessments einen besonders kritischen Punkt dar. Die diskutierten komplexen Varianzquellen in Multilevel IRT Modellen in diesem Abschnitt sollen nicht den Eindruck erwecken, dass diese Modelle in allen Situationen zu favorisieren sind. Aus diesem Kontext kann argumentiert werden, dass es keine „richtigen“ Modelle, sondern nur für den Beschreibungszweck nützliche Modelle gibt. Selbst

wenn bei hinreichend großer Stichprobe, hinreichend hoher Itemanzahl und hinreichend gutem Testdesign ein Testlet-Multilevel DIF-Modell (Abschnitt 3.5.3) dem Rasch-Modell aufgrund von Modell-Fit-Kriterien zu favorisieren ist, entscheidet letztendlich der Verwendungszweck die Modellwahl. Ist eine Teilmenge von Parametern wie Itemschwierigkeiten von Interesse, so können zuverlässige Aussagen durchaus auch (mit ggf. geringem Effizienzverlust) mit Hilfe des einfachen Rasch-Modells anstelle komplexer Modelle getroffen werden. In einer strikt formativen Interpretation der Skala (wie der Interpretation als Second Order Formative Model) argumentieren wir, dass Modell-Fit-Kriterien eine untergeordnete Rolle spielen. Im bildungspolitischen, aber auch akademischen Kontext nicht zu unterschätzen ist, dass einfache Modelle in vielen Situationen Befunde komprimierter zusammenfassen und Kommunikationsprozesse damit erleichtern. Psychometrische Modelle verschiedener Komplexität können hinsichtlich der substanziellen Interpretation ihrer Parameter je nach gewünschter analytischer Detailtiefe der Struktur des Tests im Assessment eingesetzt werden. Wir betonen, dass dieses Vorgehen nicht nach dem „richtigen“ oder dem „besten“ Modell sucht. Dessen Existenz scheint in einer Population von Schülern, Items und Testsituationen nicht zwingend gesichert.

Besonders pointiert wird die Modellwahl in einem Editorial eines kürzlich erschienen methodischen Themenhefts der Zeitschrift Developmental Psychology dargestellt (Foster & Kalil, 2008, S. 302):

> Another theme we observed among the set of submissions is that developments in methodology can run the risk of being driven by a narrow range of software, notable Mplus [...] Perhaps this relationship is being to be expected and an economist would hardly fault a manufacturer for being responsive to consumer demands. Nonetheless, when an entire field relies on a single piece of software, the field becomes subject to needs and interests of the software developers. One problem is that most developmentalists do not work with software – like Stata or R – that can be easily extended to incorporate new methods.

Die beschriebene Situation scheint auch auf Teilbereiche der empirischen Bildungsforschung übertragbar zu sein. Die Dominanz und Präferenz gewisser Softwareprodukte läuft Gefahr in einem gewissen Ausmaß die Modellwahl einzuschränken. Dies halten wir für eine äußerst unglückliche Entwicklung, da theoretisch-inhaltliche Überlegungen die Modellwahl und erst nachgelagert die Wahl einer geeigneten Software implizieren sollte.

Kapitel 4

Bedeutung der Itemauswahl und der Modellwahl für die längsschnittliche Erfassung von Kompetenzen

4.1 Einleitung

In einer Testsituation interagieren Testteilnehmerinnen und Testteilnehmer mit Items. Beide Komponenten der Testung tragen zum Zustandekommen von Testleistungen bei. Soll eine Leistungsentwicklung in einem Test im Längsschnitt untersucht werden, so muss das Zusammenwirken von Testteilnehmern und Items in der Lage sein, diese Veränderung quantitativ (zumindest im Mittel) abzubilden. Sowohl die Auswahl einer Stichprobe von Testteilnehmern als auch die konkret für den Test ausgewählten Items können dabei Variabilität in Effektgrößen der Veränderung verursachen.

In diesem Beitrag wird insbesondere der durch die Itemauswahl (auch als Item Sampling oder Itemspezifität bezeichnet) bedingte Einfluss für längsschnittliche Fragestellungen diskutiert und in den Kontext der Generalisierbarkeit von Ergebnissen aus Tests eingebettet. Die empirische Illustration erfolgt dabei anhand eines standardisierten Lesekompetenztestes für Grundschüler.

Im kommenden Abschnitt wird dafür zunächst auf einige Hauptbefunde zur Lesekompetenzentwicklung eingegangen. Danach werden im Konzept der Generalisierbarkeit von Tests Item Sampling und Model Uncertainty begrifflich eingeführt und statistische Verfahren diskutiert, die durch Item Sampling und Model Uncertainty bedingte Variabilität in statistischen Parametern quantifizieren. Die Überlegungen münden in einer kritischen Betrachtung der Modellselektionskriterien für Item-Response-Modelle, die üblicherweise für die Erfassung von Veränderung eingesetzt werden. Der Beitrag schließt mit Empfehlungen für die Publikationspraxis von zusätzlichen Fehlerquellen, die nicht mit einer Stichprobenziehung von Personen verbunden sind.

4.2 Lesekompetenz und Lesekompetenzentwicklung

Kompetente Leser verfügen über ein komplexes Bündel an Teilfähigkeiten, die in vielfältiger Weise miteinander interagieren (Artelt & Dörfler, 2010; Richter & Christmann, 2002). Die Entwicklung der Lesekompetenz beginnt bereits im Vorschulalter (Schneider & Stefanek, 2004) und setzt sich kontinuierlich über das Leben fort (Afflerbach & Paris, 2008). Am Ende der Grundschulzeit ist ein wesentlicher Entwicklungsschritt abgeschlossen, da Schüler nun zumeist in der Lage sind, einen ihnen unbekannten Text zu lesen und zu verstehen (vgl. Klicpera & Gasteiger-Klicpera, 1993; Scheerer-Neumann, 1997). Hierzu müssen Schüler in der Lage sein, die Bedeutung geschriebener Texte zu konstruieren - ein Prozess, der als Textverstehen definiert wird (Duke & Carlisle, 2011).

Aus einer Entwicklungsperspektive betrachtet sind die größten intraindividuellen Zuwächse hinsichtlich der Lesekompetenz von Schülern folglich zwischen dem Vorschulalter und dem Ende der Grundschulzeit zu beobachten. Die Entwicklung in diesem Zeitraum kann am treffendsten mit einer negativ beschleunigten Wachstumskurve beschrieben werden: Besonders hohe Zuwächse im Bereich Lesekompetenz/Textverstehen sind zwischen dem Kindergarten und der ersten Klassenstufe (mittleres $d = 1.52$) sowie zwischen der ersten und zweiten Klassenstufe (mittleres $d = .97$) zu verzeichnen, während sich der Kompetenzzuwachs dann zwischen den Klassenstufen zunehmend abflacht (Hill, Bloom, Black & Lipsey, 2008). Hill und Kollegen (2008) berichten Effektstärken der Veränderung zwischen der dritten und vierten Klassenstufe von .24 bis .54 bezogen auf sieben unterschiedliche Lesetests (mittleres $d = .36$). Diese verschiedenen Befunde können einerseits durch unterschiedliche Repräsentationen von unterschiedlichen Konstrukten der Lesekompetenz in diesen Tests verursacht sein. Andererseits trägt allerdings auch die Art der Testaufgaben im Sinne des Item Sampling wesentlich zur Effektstärke der Veränderung bei. Für deutsche Schüler wurde in querschnittlich angelegten Untersuchungen ein Leistungsunterschied zwischen Dritt- und Viertklässlern für die Lesekompetenz von etwa $d = .6$ sowohl in nationalen Erhebungen der Bildungsstandards als auch in der internationalen Vergleichsstudie IGLU ermittelt (Böhme & Robitzsch, 2009).

4.3 Konzept der Generalisierbarkeit für Tests

In der Generalisierbarkeitstheorie (Brennan, 2001a) werden verschiedene Facetten definiert, für die eine Generalisierung von Ergebnissen notwendig erscheinen. Soll in Tests nicht nur auf die konkrete im Test eingesetzte Itemmenge, sondern auf ein potenzielles Universum von Items eines Leistungsbereiches generalisiert werden, so muss die Variationsquelle der Itemauswahl (so genanntes *Item Sampling*) in Rechnung gestellt werden. Zusätzlich können aufgrund der konkreten Wahl statistischer Modelle Modellergebnisse variieren (*Model Uncertainty*; Clyde & George, 2004; Young, 2009). Dabei können verschiedene Verfahren der Skalierung oder verschiedene Methoden zur Verlinkung von Tests zu unterschiedlichen Messzeitpunkten mögliche Analysemodelle darstellen, die jeweils zu verschiedenen Ergebnissen führen können. Die Kombination mehrerer Modelle in einer statistischen Inferenz (*Multi Model Inference*) findet schon seit einiger Zeit in der Literatur Berücksichtigung (Burnham & Anderson, 2002; Draper, 1995; Montgomery & Nyhan,

2010).

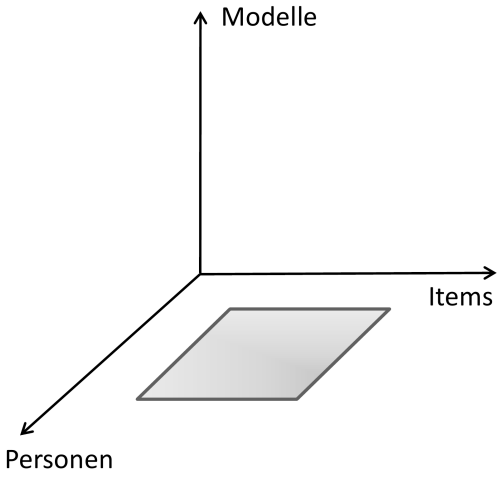

Abbildung 4.1: *Generalisierbarkeit für Tests. Dargestellt sind die Facetten Personen, Items und Modelle.*

In Abbildung 4.1 werden die drei Facetten der Generalisierbarkeit von Parametern für Tests unterschieden: die Wahl von Personen, Items und Modellen. Die Durchführung einer statistischen Inferenz für einen interessierenden Parameter setzt die Annahme von Wahrscheinlichkeitsverteilungen voraus. Diese Verteilungen in der Generalisierung von einer konkreten Stichprobe auf eine wohl definierte oder abstrakte hypothetische Population existieren nicht zwingend aufgrund einer klar definierten Stichprobenziehung, sondern werden vom Forscher in Abhängigkeit einer interessierenden Inferenz spezifiziert (Kane, 2011; Kass, 2011). Statistische Inferenz kann sich dabei auf eine Stichprobenziehung (Sampling) im eigentlichen Sinne – der so genannten designbasierten Inferenz – beziehen. Davon ist die modellbasierte Inferenz abzugrenzen, bei der der Schritt der Generalisierung nicht aufgrund eines konkreten Samplings, sondern aufgrund der Spezifikation eines datengenerierenden Modells entsteht (Särndal, Swensson & Wretman, 1992).

Unter *Person Sampling* versteht man die konkrete oder hypothetische Stichprobenziehung von Personen, unter *Item Sampling* die konkrete oder hypothetische Stichprobenziehung von Items und unter *Model Inference* (auch *Model Sampling*) die Definition einer Inferenz oder einer konkreten Modellauswahl aus einer Menge endlich oder unendlich vieler theoretisch plausibler statistischer Modelle. Bei Vorliegen einer Fragestellung sollte dabei der Forscher vor Durchführung der statistischen Analyse eine möglichst kleine Menge theoretisch plausibler Modelle wählen (Burnham & Anderson, 2002).

Variation in interessierenden statistischen Parametern kann dabei in Abhängigkeit des Samplings in diesen drei Facetten entstehen, so dass wir für die Stichprobenvarianz einer allgemeinen Parameterschätzung $\hat{d}$ (z. B. einem Mittelwert oder einer Effektstärke) näherungsweise formulieren können:

$$Var(\hat{d})_{Total} = Var(\hat{d})_{Personen} + Var(\hat{d})_{Items} + Var(\hat{d})_{Modelle} \tag{4.1}$$

Hierbei sind zunächst Interaktionen von Fehlerquellen ignoriert worden, so dass wir von

Unabhängigkeit der drei Facetten ausgehen (für mögliche Konsequenzen dieser Annahme sei auf die Diskussion verwiesen). Die Wurzel $\sqrt{Var(\hat{d})_{Total}}$ ist der Standardfehler der Schätzung $\hat{d}$, der sich aus der Unsicherheit der Inferenz für Personen, Items und Modellen zusammensetzt. In typischen Anwendungen der sozialwissenschaftlichen Forschung werden meistens nur durch Person Sampling (d.h. durch Ziehung von Personenstichproben) verursachte Standardfehler berichtet, so dass für inferenzstatistische Verfahren oft fälschlicherweise $Var(\hat{d})_{Total} = Var(\hat{d})_{Personen}$ angenommen wird. Demzufolge wird die durch Item Sampling bedingte Variabilität (siehe im Kontext von Large-Scale Assessments wie der PISA-Studie: Wu, 2010) und die durch Modellwahl bedingte Unsicherheit ignoriert (Berk, Brown & Zhao, 2010; Chatfield, 1995), so dass Standardfehler unterschätzt und statistische Tests zu liberal durchgeführt werden.

Der vorliegende Beitrag fokussiert den Einfluss von Item Sampling und Model Uncertainty für die Interpretation der Lesekompetenzentwicklung.

4.3.1 Item Sampling

Im Kontext von Tests findet das Konzept einer unendlichen Itempopulation schon seit Längerem Berücksichtigung (Husek & Sirotnik, 1967; Lord & Novick, 1968). Im Hinblick auf längsschnittliche Analysen weisen Hutchison (2008), Michaelides (2010) sowie Michaelides und Haertel (2004) darauf hin, dass die Wahl von Linkitems zwischen mehreren Studien in Längsschnittanalysen die Interpretation eines Item Sampling erhalten soll. Ist die Anzahl dieser Linkitems zu gering oder die konkrete Auswahl der Linkitems nicht repräsentativ für die gesamte Itemmenge, so können verzerrte Schätzungen für Leistungsentwicklungen resultieren (Mazzeo & von Davier, 2008; van den Heuvel-Panhuizen, Robitzsch, Treffers & Köller, 2009). In der psychometrischen Literatur zur Faktorenanalyse stellt das so genannte *Domain Sampling* oder der *Psychometric Inference* die Inferenz auf eine Itempopulation dar (Cronbach & Shavelson, 2004, Kaiser & Caffrey, 1965; McDonald, 2003; McDonald & Mulaik, 1979; Lord, 1955; Tryon, 1957). Historisch gesehen wurden mit dem Aufkommen der Item-Response-Modelle in den meisten Anwendungen Items als feste Facette (also *Fixed Items*) interpretiert, d.h. bei bekannten Itemparametern spielt in diesen Modellen die konkrete Auswahl von Items für Aussagen über Fähigkeiten der Schüler keine Rolle, was man in der Literatur mitunter als Folge der Eigenschaft der so genannten spezifischen Objektivität des Rasch-Modells ansieht (Rasch, 1977).

Implikationen der Betrachtung der Behandlung von Items mit Zufallseffekten (*Random Items*) diskutieren De Boeck (2008) sowie Briggs und Wilson (2007) in einer Verbindung von Item-Response-Modellen und der Generalisierbarkeitstheorie. De Boeck (2008) argumentiert konträr zur vorzufindenden Praxis, dass für diagnostische Zwecke eher Personen als feste Facette und Items als zufällige Facette zu interpretieren seien, da diagnostische Aussagen für jede einzelne Person separat und nicht für eine gesamte Population (oder Subpopulationen) getroffen werden[1]. Brennan (2011) merkt jedoch die strikten

[1] In einer „klassischen Perspektive“ des Rasch-Modells stellen Itemschwierigkeit und Personenfähigkeiten feste, unbekannte Größen dar. Eine Schätzung der Verteilung dieser Parameter scheint dabei nicht von primärem Interesse, so dass Fixed Items und Fixed Persons resultieren (Kubinger, Rasch & Yanagida, 2011, S. 555 ff.).

Voraussetzungen unter der Fixed Items Perspektive in IRT-Modellen im Gegensatz zur Generalisierbarkeitstheorie an, die die Auswahl von Items explizit als Variabilitätsquelle thematisiert. Auch einige theoretische Arbeiten zur Definition latenter Variablen in Item-Response-Modellen gehen von einer unendlichen Itempopulation aus (Cliff & Donoghue, 1992; Douglas, 1997, 2001; Junker, 1993; Stout, 1990). Aus diesen IRT-Modellen abgeleitete Parameter können demzufolge mit maximaler Präzision nur bei einer unendlich großen Personenstichprobe und einer unendlich großen Itemstichprobe erhalten werden. Modellgeltungstests für IRT-Modelle, in denen die konkret im Test eingesetzten Items als feste Facette behandelt werden, würden dann das Item Sampling ignorieren und könnten zu falschen Schlussfolgerungen gelangen.

Für die Erfassung längsschnittlicher Leistungsentwicklungen ist jüngst die Methode des Item Samplings wieder aufgegriffen worden (Strathmann & Klauer, 2010). Strathmann und Klauer weisen darauf hin, dass bei der Generierung von Tests auf eine hinreichende Repräsentation bestimmter Komponenten des zu erfassenden Konstrukts (Subkompetenzen) und (ggf. konstruktirrelevanten) Oberflächenmerkmalen von Items (Stimuli, Situationen, Itemformate) zu achten sei, so dass die Ziehung von Items dann aus einer stratifizierten Itemstichprobe erfolgt (siehe Tryon, 1957). Aus der Survey-Statistik ist bekannt, dass im Allgemeinen eine Stratifikation für Stichproben die Präzision der Schätzungen von Populationsparametern erhöht (Särndal et al., 1992). Für mehrdimensionale Konstrukte können in einem stratifizierten Item Sampling für einen komplex zusammengesetzten Test die einzelnen Strata der Itempopulation Subdimensionen entsprechen.

Statistische Inferenz für Itemstichproben kann relativ einfach mit statistischen Resampling-Verfahren (Cameron & Trivedi, 2005) durchgeführt werden. Gibt es in einem Test beispielsweise $I = 20$ Items, so wird bei der Methode des Jackknife eine bestimmte Analyse (etwa die Berechnung einer Effektstärke) unter Ausschluss von jeweils einem Item durchgeführt, so dass $I = 20$ Analyseergebnisse resultieren. Die Variation zwischen den verschiedenen Ergebnissen kann für die Berechnung eines durch Item Sampling verursachten Standardfehlers verwendet werden (siehe für eine Anwendung im Kontext der Verlinkung von Items zwischen mehreren Studien Monseur & Berezner, 2007). Es bezeichne d eine Statistik, die aus der Einbeziehung aller Items entsteht, sowie $d^{(-i)}$ diese Statistik, wenn das i-te Item aus der Berechnung entfernt wurde. Mittels Jackknife ergibt sich dann bei I Items die Berechnung des Standardfehlers gemäß

$$SE(d) = \sqrt{\frac{I-1}{I} \cdot \sum_{i=1}^{I} \left(d^{(-i)} - d\right)^2} \tag{4.2}$$

Alternativ können zufällige Itemeffekte auch mittels Varianzkomponentenmodellen (De Boeck, 2008; van den Berg et al., 2007) oder Generalized Linear Mixed Models (Baayen, Davidson & Bates, 2008; Doran, Bates, Bliese & Dowling, 2007) spezifiziert werden. Item Sampling im Kontext cross-nationaler Vergleiche diskutieren de Jong et al. (2007).

Eine integrierte Betrachtung von Person und Item Sampling schlagen Xu und von Davier (2010) mittels der Anwendung von Jackknife unter gleichzeitiger Elimination von jeweils einer Person (oder einer Personengruppe) bzw. eines Items vor (sog. *Double Jackknife*), so dass dann auch Interaktionen der Fehlerquellen der beiden Facetten Personen

und Items zugelassen sind. Das Prinzip des Jackknife kann auch als Verfahren zur Untersuchung der Stabilität von Modellergebnissen unter Auswahl von Teildatensätzen (Personen und/oder Items) angesehen werden, ohne konkrete Annahmen eines datengenerierenden Modells oder die Ziehung einer Stichprobe treffen zu müssen (sog. *Stability under Data Selection*; de Leeuw, 1988; Gifi, 1990).

4.3.2 Zur Wahl eines IRT-Modells

Für die Beschreibung längsschnittlicher Veränderungen können verschiedene Skalierungsvorschriften für vorliegende Tests gewählt werden. Aussagen zu Effektgrößen der Veränderung hängen von der durch die Skalierung gewonnenen Metrik ab. In der Literatur wird allerdings mitunter argumentiert, dass aus einem IRT-Modell abgeleitete Personenschätzer eher als ordinalskaliert und nicht als intervallskaliert anzusehen sind (Cliff & Keats, 2003; Lord, 1980; Ramsay, 1996; Stout, 1990; Zwick, 1992). In diesem Fall wäre jede monotone Skalentransformation zulässig und nur auf Rangdaten basierende Aussagen sinnvoll interpretierbar. Verfechter des Rasch-Modells argumentieren dagegen, dass nur das Rasch-Modell als einziges IRT-Modell Personenschätzungen auf Intervallskalenniveau generiere (Fischer, 1995a) und verwenden dabei dessen Eigenschaft der so genannten spezifischen Objektivität (Rasch, 1977). Diese Eigenschaft beruht darauf, dass individuelle Lösungswahrscheinlichkeiten $P(X_{pi} = 1)$ für Items als Transformation einer Differenz aus Personenfähigkeiten und Itemschwierigkeiten parametrisiert werden (Steyer & Eid, 2001). Stellt man diese Eigenschaft allerdings als alleinige Forderung in dieser Modellklasse, so wird die Herausstellung des Rasch-Modells in der Klasse der IRT-Modelle relativiert, da jede streng monotone Funktion g (die so genannte Linkfunktion, die die Differenz aus Personenfähigkeit und Itemschwierigkeit in eine Wahrscheinlichkeit transformiert) mit $P(X_{pi} = 1) = g(\theta_p - b_i)$ ein IRT-Modell definiert, das spezifisch objektiv ist (Goldstein, 1980; Goldstein & Wood, 1989; McDonald, 1999, S. 291).

In der statistischen Literatur werden Methoden vorgeschlagen, die die Linkfunktion g neben Item- und Personenparametern durch eine geeignete Parametrisierung als Abweichung zur logistischen Funktion schätzen (Stukel, 1988). Unter der Bedingung, dass für die Items keine Parameter für Rateeffekte oder Flüchtigkeitsfehler definiert werden, zeigt Peress (2012) unter der Annahme unendlich vieler Personen und unendlich vieler Items die Identifizierbarkeit dieses IRT-Modells (d.h. dessen Schätzbarkeit) unter der Fixed Items und Fixed Persons Perspektive.

Bei endlich vielen Items und endlich vielen Personen kann die Wahl einer Linkfunktion g als Model Uncertainty angesehen werden, wenn aus substanzwissenschaftlichen Gründen keine Linkfunktion zu präferieren ist. Die Bedeutung des Rasch-Modells wird beispielsweise durch Irtel (1995a) oder Scheiblechner (2007) relativiert, die die Eigenschaft der spezifischen Objektivität unter schwächeren Annahmen in allgemeineren Item-Response-Modellen zeigen. Wendet man allerdings auf ein IRT-Modell mit einer Linkfunktion g anstelle einer Differenz eine Parametrisierung von Wahrscheinlichkeiten als Quotient der Form $P(X_{pi} = 1) = g(\theta_p/b_i)$ dieselbe Argumentation an, so würden sich für die IRT-Scores sogar verhältnisskalierte Variablen ergeben. Ein von Ramsay (1989; siehe auch Wainer, 2010a) vorgeschlagenes Quotienten-Modell kann aufgrund kognitionspsychologischer Überlegungen abgeleitet werden (van der Maas, Molenaar, Maris, Kievit & Bors-

boom, 2011) .

Schulz und Nicewander (1997) vergleichen Scores auf der Rohscore-Metrik und verschiedene IRT-Metriken für Längsschnittdaten und kommen zum Schluss, dass die Metrik nur aus substanziellen und nicht aus statistischen Gründen gewählt und eher als ordinalskaliert angesehen werden kann. Nur Sensitivitätsanalysen (Morgan & Winship, 2007) können dann Aussagen darüber treffen, in welchem Ausmaß Effektgrößen von einer konkreten Wahl der Metrik abhängen (Zwick, 1992). In Sensitivitätsanalysen werden dabei mehrere Effektgrößen unter verschiedenen Varianten der Itemselektion und der Skalierung berechnet.

4.3.3 Modellselektion, (Multi) Model Inference, Model Averaging oder Model Sampling?

„Konventionelle statistische Inferenz“ geht davon aus, dass ein datengenerierendes Modell korrekt spezifiziert ist und Standardfehlerschätzungen für einen interessierenden Parameter gegeben das gewählte Modell berechnet werden. Findet vor dieser Inferenz eine Modellselektion auf Grund empirischer Kriterien (wie beispielsweise Fitstatistiken oder Informationskriterien) statt, so bleibt diese Unsicherheit der Modellselektion in „konventioneller Inferenz“ unberücksichtigt (Berk et al., 2010; Chatfield, 1995). Die Multi Model Inference integriert den Schritt der Modellwahl in eine statistische Inferenz, die nicht von der Korrektheit und der a priori Kenntnis eines Modells ausgeht (Draper, Hodges, Leamer, Morris & Rubin, 1987; Burnham & Anderson, 2002). Dabei muss der Forscher zunächst eine Menge theoretisch plausibler Modelle spezifizieren (siehe hierfür Weigel, Knutti, Liniger & Appenzeller, 2010, S. 4178) und jedes der Modelle mit einer Priorwahrscheinlichkeit (bzw. einem Modellgewicht) versehen (Draper, 1995), die den Glaubensgrad des Forschers in das jeweilige Modell ausdrückt (siehe Garthwaite & Mubwandarikwa 2010 zur Definition dieser Gewichte durch den Forscher). Die Modellgewichte werden dann beim Verfahren des Model Averaging (Brock, Durlauf & West, 2003) bei der Aggregation von Paramaterschätzungen aus allen Modellen verwendet (Clemen & Winkler, 1999; Levin & Williams, 2003) oder in einem Bayesianischen Ansatzes mittels Posteriorwahrscheinlichkeiten, die den Fit des Modells gegeben die Daten widerspiegeln, revidiert (sog. *Bayesian Model Averaging*; Hoeting, Madigan, Raftery & Volinsky, 1999). In der Literatur werden jedoch auch Methoden vorgeschlagen, die Modellgewichte beim Model Averaging so bestimmen, dass die Variabilität der interessierenden Parameterschätzung (den RMSE) minimiert wird (Claeskens & Hjort, 2003; Longford, 2003, 2008; Weigel et al., 2010). Bezeichnen w_m Modellgewichte in Modellen $m = 1, \ldots, M$ (wobei $\sum_{m=1}^{M} w_m = 1$) und δ_m eine Parameterschätzung im m-ten Modell, dann berechnet sich im Model Averaging ein kombinierter Schätzer δ als

$$\delta = \sum_{m=1}^{M} w_m \delta_m \tag{4.3}$$

Die Varianz dieser Schätzung ist als Summe der Varianzen der Schätzungen innerhalb von Modellen und der Varianz zwischen Modellen gegeben (Levin & Williams, 2003):

$$Var(\delta) = \sum_{m=1}^{M} w_m Var(\delta_m) + \sum_{m=1}^{M} w_m \left(\delta_m - \delta\right)^2 \tag{4.4}$$

Variieren die Parameterschätzungen also zwischen Modellen, so ist eine zusätzliche Varianzquelle für die Facette der Modellwahl zu berücksichtigen[2]. Wenn die Modellgewichte anhand des gleichen Datensatzes bestimmt werden, werden in der Literatur alternative Bestimmungen der Varianz vorgeschlagen (Burnham & Anderson, 2002).

Gemäß dieses Typs statistischer Inferenz würden Modelle denselben Status wie Personen und Items im Konzept der Generalisierbarkeit von Tests erlangen (Brock, Durlauf & West, 2007). Ein Model Sampling ist dabei als eine „theoretische Stichprobenziehung" von Modellen gemäß a priori Gewichten oder Posteriorwahrscheinlichkeiten und nicht zwangsläufig als eine Zufallsziehung aus einer Population von Modellen zu interpretieren (Kane, 2002; Kass, 2011).

4.3.4 Effektgrößen für Längsschnittdaten

Für die Beschreibung mittlerer Veränderung werden in der Literatur verschiedene Effektgrößen vorgeschlagen (Algina, Keselman & Penfield, 2005). Eine Möglichkeit ist die Standardisierung der Differenz aus den Mittelwerten des zweiten und ersten Messzeitpunktes an der Standardabweichung des ersten Messzeitpunktes. Diese Effektgröße d misst einen mittleren Unterschied bezüglich der zu T1 vorliegenden Streuung s_1:

$$d = \frac{\overline{x}_2 - \overline{x}_1}{s_1} \tag{4.5}$$

Alternativ kann der Mittelwertunterschied auch an der mittleren Standardabweichung von T1 und T2 standardisiert werden, so dass die Effektgröße d^* entsteht:

$$d^* = \frac{\overline{x}_2 - \overline{x}_1}{\sqrt{\left(s_1^2 + s_2^2\right)/2}} \tag{4.6}$$

Fällt in Abhängigkeit der Skalierungsmethode (d.h. der Wahl eines bestimmten IRT-Modells) die Standardabweichung s_2 zu T2 deutlich verschieden aus, dann verringert d^* im Vergleich zu d die Abhängigkeit der Effektgröße von der Wahl der Skalierungsmethode, da die Standardabweichung s_2 zu T2 in d^* im Gegensatz zu d berücksichtigt wird (Briggs & Weeks, 2009). Besonders große Abweichungen können dabei bei Skalierungen mit einem einparametrischen logistischen (Rasch-)Modell im Vergleich zu einem dreiparametrischen logistischen Modell entstehen. Besitzen die aus einem IRT-Modell gewonnenen Messwerte nur Ordinalskalenniveau, so sind nur auf Rangplatzordnungen definierte

[2]Dabei ist die Definition der Unsicherheit der Inferenz über mehrere Modelle nicht offensichtlich. Buckland, Burnham und Augustin (1997, S. 204) schreiben: „How to estimate the variance of δ depends on our philosophy." Damit kann gemeint sein, ob Abhängigkeiten zwischen verschiedenen Modellergebnissen mitmodelliert werden sollen oder von unabhängigen „Modellziehungen" ausgegangen werden kann (siehe die anschließende Diskussion in diesem Abschnitt).

Effektgrößen heranzuziehen, die die Invarianz unter streng monotonen Transformationen der IRT-Metrik sichern. Die nichtparametrische Effektgröße V^* transformiert die Wahrscheinlichkeit $P(X_2 \geq X_1)$, so dass Scores X_2 zu T2 mindestens so groß wie Scores X_1 zu T1 sind, nichtlinear (Ho, 2009; Zwick, 1992):

$$V^* = \sqrt{2} \cdot \Phi^{-1} \left[P(X_2 \geq X_1) \right] \tag{4.7}$$

Dabei bezeichnet Φ^{-1} die inverse Standardnormalverteilungsfunktion, die Wahrscheinlichkeiten den Quantilen der Standardnormalverteilung zuordnet. Sind die Scores zu T1 und T2 normalverteilt, so stimmen V^* und d^* überein. Für diskrete Scores mit vielen Bindungen (d. h. gleichen Scores zu T1 und T2), wird in diesem Artikel zusätzlich die Effektgröße V^{**} vorgeschlagen, die gleiche Rangplätze zu T1 und T2 in gleichem Ausmaß auf beide Zeitpunkte aufteilt:

$$V^{**} = \sqrt{2} \cdot \Phi^{-1} \left[P(X_2 > X_1) + \frac{1}{2} \cdot P(X_2 = X_1) \right] \tag{4.8}$$

Bei einer großen Anzahl von Bindungen kann das Intervall $[V^{**}, V^*]$ auch einen Bereich „plausibler Effektgrößen“ darstellen, wenn Unsicherheit in der statistischen Behandlung von Bindungen vorliegt. Die Auswahl der Effektgröße könnte dabei allerdings wie die Modellwahl als eine weitere Facette angesehen werden, für die ein Forscher Priorwahrscheinlichkeiten (d.h. Gewichte für jede der Effektgrößen) spezifizieren kann. Dies soll aber in diesem Beitrag nicht verfolgt werden.

4.3.5 Item Parameter Drift

Die Beschreibung längsschnittlicher Veränderung mit Hilfe von IRT-Modellen macht die Setzung von Restriktionen für Itemparameter notwendig. Häufig wird ein invariantes Itemfunktionieren für alle Messzeitpunkte für die Anwendung von IRT-Modellen vorausgesetzt (Rupp & Zumbo, 2006). Werden beispielsweise in einem zweidimensionalen Rasch-Modell Kompetenzen zu den zwei Zeitpunkten T1 und T2 als zwei Dimensionen spezifiziert (Fischer, 1995b), so müssen unter einer Invarianzannahme alle Items den gleichen Fähigkeitsunterschied zu den beiden Zeitpunkten homogen messen. Kontrolliert man Fähigkeitsunterschiede zu T1 und T2, so wird eine von Null verschiedene Differenz zweier Itemschwierigkeiten zu T1 und T2 für ein Item als *Item Parameter Drift* (IPD), einer speziellen Form differenziellen Itemfunktionierens, bezeichnet (Holland & Wainer, 1993). Es wird in der Literatur häufig empfohlen, Items mit IPD für die Skalierung der Längsschnittdaten zu entfernen (Fischer, 1995b), da diese Items die Messpräzision verringern könnten. Diese Perspektive trifft jedoch nur auf den Fall von Fixed Items zu. Unter der Perspektive des Item Sampling kann aber jedes Items zu den beiden Zeitpunkten differenziell funktionieren und mit einer größeren Itemanzahl kann die Schätzgenauigkeit erhöht werden. Dieser Sachverhalt lässt sich formal präzisieren: Ist σ^2_{IPD} die Varianz des Item Parameter Drifts für die Itempopulation, dann ergibt sich für den durch Item Sampling (und damit IPD) bedingten Standardfehler der mittleren Veränderung $\sigma_{IPD}/\sqrt{I}$, wenn I die Anzahl der bei beiden Zeitpunkten zugleich eingesetzten Items bezeichnet. Prinzipiell lässt sich die mittlere Veränderung auf der durch das Rasch-Modell definierten Metrik

als Mittelwert der mit allen Items gemessenen Veränderung auffassen. Ein Entfernen von Items aus dieser Mittelwertsbildung kann nur dann effizient sein, wenn es sich bei diesen Beobachtungen um Ausreißerwerte handelt (siehe auch entsprechende Überlegungen aus der robusten Statistik für die Schätzung eines Populationsmittelwertes mit getrimmten Mittelwerten; Maronna, Martin & Yohai, 2006). Eine Schätzung der mittleren Veränderung kann bei einem Entfernen von Items mit betraglich großem Item Parameter Drift relativ konstant bleiben, während dessen sich der Standardfehler auf Grund von Item Sampling infolge der geringeren Itemanzahl erhöht.

4.4 Fragestellungen

In dieser Arbeit soll anhand des ELFE-Tests festgestellt werden, ob Item Sampling eine Bedeutung für die Erfassung längsschnittlicher Veränderung von Lesekompetenz besitzt. Die durch diese Variabilitätsquelle verursachten Standardfehler sollen mit den durch Person Sampling verursachten Standardfehlern verglichen werden. Außerdem soll auch die mit der Wahl der Metrik einhergehende Unsicherheit in der Berechnung der Effektgrößen berücksichtigt werden, um die Stabilität von Aussagen über Lesekompetenzentwicklung in der Primarstufe hinsichtlich der Model Uncertainty zu beurteilen. Die Befunde sollen ermutigen, in Forschungskontexten andere Quellen als das Person Sampling als zusätzliche Grundlage statistischer Inferenz einzubeziehen.

4.5 Methode

4.5.1 Stichprobe

Die vorliegenden Analysen stützen sich auf Daten, die im Rahmen der Teilprojekte zur Kompetenzentwicklung der Bamberger BiKS-Studie (Bildungsprozesse, Kompetenzentwicklung und Formation von Selektionsentscheidungen im Vor- und Grundschulalter) erhoben wurden. In dieser längsschnittlich angelegten Studie wurden Schüler vom Beginn des zweiten Halbjahres der dritten Klasse bayerischer und hessischer Grundschulen in halbjährlichem Abstand in verschiedenen kognitiven und nicht-kognitiven Bereichen getestet sowie umfassend zu schulischen und familiären Hintergrundvariablen befragt (Kurz, Kratzmann & von Maurice, 2007). In der vorliegenden Arbeit werden drei Messzeitpunkte aus der Grundschule analysiert: Messzeitpunkt T1 stammt dabei aus dem zweiten Halbjahr der dritten Klassenstufe, die Messzeitpunkte T2 bzw. T3 aus dem ersten bzw. zweiten Halbjahr der vierten Klassenstufe. Zu mindestens einem der drei Messzeitpunkte liegen Messungen der Lesekompetenz von N = 2380 Schülern aus insgesamt 154 Klassen vor. Die Schüler sind zu Beginn der Untersuchung (T1) 9.3 Jahre alt (SD = .47). Beide Geschlechter sind zu gleichen Teilen in der Stichprobe vertreten (49.9 % Jungen).

4.5.2 Instrumente

Zur Erfassung der Lesekompetenz lasen die Schüler 13 unterschiedlich lange Texte und bearbeiteten anschließend ein Set von Multiple-Choice-Items zu den jeweiligen Texten. Die

Lesekompetenz wurde zu drei Messzeitpunkten mit 20 Items aus dem Subtest Textverständnis des ELFE 1-6 (Lenhard & Schneider, 2005) gemessen. Die internen Konsistenzen des Instruments waren zu allen Messzeitpunkten zufriedenstellend ($\alpha_{T1} = .86$, $\alpha_{T2} = .86$ und $\alpha_{T3} = .86$). Für den ersten Messzeitpunkt wurde zur Überprüfung der Faktorenstruktur zusätzlich eine auf der tetrachorischen Korrelationsmatrix basierende exploratorische Faktorenanalyse in Mplus 6.0 (Muthén & Muthén, 1998-2010) gerechnet. Der erste Faktor klärt danach ca. 45% der Varianz der Daten auf. Das Eigenwerteverhältnis zwischen erstem und zweitem Faktor liegt bei 2.14. Dies wird als hinreichender Indikator zur Feststellung der Eindimensionalität der Daten gewertet (Hattie, 1985).

4.5.3 Behandlung fehlender Daten

Von den N=2380 Schülern liegen nur von 1855 Schülern Messungen der Lesekompetenz zu allen drei Zeitpunkten vor. Der Ausfall nahm von T1 zu T3 zu (T1: 4.6%, T2: 8.4% sowie T3: 15.3%). Die Teilnahme am Test zu T2 (bzw. T3) hängt dabei relativ stark von der Leistung zu T1 ab ($d = .33$ bzw. $d = .32$). Demzufolge ist nicht von Missing Completely at Random (MCAR) im Datensatz auszugehen. Bei der Behandlung fehlender Daten, die nicht MCAR sind, kann man grundlegend zwischen modellbasierten Verfahren, bei denen alle beobachteten Daten in die Analyse eingehen und die Missings modellimplizit behandeln, und imputationsbasierten Verfahren, die fehlende Daten ersetzen und somit modellexplizit behandeln, unterscheiden (Lüdtke & Robitzsch, 2010). In diesem Beitrag werden beide Methoden eingesetzt. Nur für ein später verwendetes IRT-Modell, bei dem die Items zu den Zeitpunkten separat skaliert werden, wird dabei auf imputierte Daten zurückgegriffen, um verzerrte Parameterschätzungen zu vermeiden. In das Imputationsmodell gehen 60 Items (20 Items zu den 3 Zeitpunkten) ein. Als Vorhersagemodell für jedes Item dient eine lineare Regression mit allen anderen Items als Prädiktoren, wobei die Methode des Predictive Mean Matchings dichotome Imputationen der Item Scores ermöglicht (Münnich, 2005). Zur Verringerung von Multikollinearitäten im Imputationsmodell wurden in jedem Iterationsschritt für jedes zu imputierende Item die jeweils übrigen 20+20+19=59 Items als Prädiktoren im linearen Regressionsmodell mit der Methode der Partial Least Squares Regression (Mevik & Wehrens, 2007) auf 10 unkorrelierte Prädiktoren unter praktischer Konstanthaltung des R-Quadrats reduziert. Das Imputationsmodell bildet dabei komplexere Beziehungen im Vergleich zu den nachfolgend spezifizierten IRT-Modellen ab. Insgesamt wurden 20 imputierte Datensätze abgespeichert. Das R-Paket mice wird zur Imputation verwendet (van Buuren & Groothuis-Oudshoorn, 2011).

4.5.4 Statistische Analysen

Für die im Abschnitt „Effektgrößen für Längsschnittdaten“ eingeführten Parameter auf der Basis von Rohwerten (Summenscores) liegen nicht für alle Größen bekannte Stichprobenverteilungen für die statistische Inferenz vor, so dass Resampling-Verfahren zum Einsatz kommen (Cameron & Trivedi, 2005). Die Bestimmung von Standardfehlern aufgrund von Person Sampling erfolgt mit der Bootstrap-Methode, bei der Schüler aus der realisierten Stichprobe wiederholt mit Zurücklegen gezogen werden. Da Schüler innerhalb von Klassen geschachtelt sind, wird bei diesem Vorgehen allerdings die Clusterstruktur in

den Daten ignoriert. In einem zweiten Verfahren wird daher auf der Ebene der Klassen ein Jackknife-Verfahren durchgeführt, das zu Standardfehlern unter Berücksichtigung der Clusterstruktur führt (van der Leeden, Meijer & Busing, 2007). Bei einer substanziell von Null verschiedenen Intraklassenkorrelation werden daher mit dem Jackknife-Verfahren höhere Standardfehlerschätzungen als mit dem Bootstrap-Verfahren zu erwarten sein. Aufgrund von Item Sampling verursachte Standardfehler werden ebenso mittels Jackknife bestimmt, wobei immer ein Item unter Konstanthaltung der Personenstichprobe aus den Analysen entfernt wird. Interaktionen von Personen und Items finden in der Bestimmung der Standardfehler der Effektgrößen keine Berücksichtigung. Alle anderen eingesetzten Modelle operationalisieren Schülerkompetenzen mit Hilfe latenter Variablen.

In Modellgleichung (4.9) wird dabei die Wahrscheinlichkeit $P(X_{pti} = 1)$ einer für die korrekte Beantwortung des Items i zum Zeitpunkt t durch den Schüler p mit einem logistischen linearen gemischten Modells formuliert:

$$\text{logit } \{P(X_{pti} = 1)\} = \mu_t + u_p + u_{pt} + u_i + u_{it} + u_{pi} \tag{4.9}$$

Dabei ist durch μ_t der Mittelwert zum Zeitpunkt t gegeben. Alle weiteren mit u bezeichneten Variablen u_p, u_{pt}, u_i, u_{it} und u_{pi} sind zufällige Effekte und sind untereinander unkorreliert sowie besitzen homoskedastische Varianzkomponenten σ_p, σ_{pt}, σ_i, σ_{it} und σ_{pi} (wobei σ hier die Standardabweichung bezeichnet), die unter einer Normalverteilungsannahme aller u-Variablen geschätzt werden. Items und Personen stellen in Gleichung (4.9) zufällige Facetten dar und das Modell fällt in die Klasse sog. *IRT Variance Decomposition Models* (van den Berg et al., 2007). Gegenüber den in der Generalisierbarkeitstheorie (G-Theorie) verwendeten Modellen besteht nur der Unterschied, dass Gleichung (4.9) die Wahrscheinlichkeiten $P(X_{pti} = 1)$ einer logistischen Funktion unterwirft. Verwendet man stattdessen die in der G-Theorie übliche identische Linkfunktion (die auf der Ebene der Rohscores operiert), so ergibt sich anstelle von Gleichung (4.9):

$$X_{pti} = \mu_t + u_p + u_{pt} + u_i + u_{it} + u_{pi} + e_{pit} \tag{4.10}$$

Anstelle der logistischen Linkfunktion wurde dabei in Gleichung (4.10) ein normalverteiltes Residuum e_{pti} eingeführt. In diesem Modell wird dabei die Varianzkomponente σ_{pti} zusätzlich geschätzt.

Aus den Modellgleichungen (4.9) und (4.10) kann man die Effektgröße der Veränderung von T1 nach T2 unter Annahme der Homoskedastizität mit $d = d^* = \frac{\mu_2 - \mu_1}{\sqrt{\sigma_p^2 + \sigma_{pt}^2}}$ ableiten. Da die Varianzen über die Zeitpunkte hinweg in Gleichung (4.9) und (4.10) konstant sind, stimmen die Effektgrößen d und d^* überein.

Die Modelle der Gleichungen (4.9) und (4.10) erlauben primär deskriptive Beschreibungen der Größe der Varianzkomponenten der Facetten p, pt, i, it und pi. Die Annahme zeitkonstanter Varianzen und gleicher Korrelationen zwischen allen drei Zeitpunkten wird jedoch in folgendem mehrdimensionalen Rasch-Modell (4.11) abgeschwächt:

$$\text{logit } \{P(X_{pti} = 1)\} = \mu_t + \theta_{pt} - b_i \tag{4.11}$$

Dabei ist der Fähigkeitsvektor $(\theta_{p1}, \theta_{p2}, \theta_{p3})$ dreidimensional zentriert normalverteilt. Die Itemschwierigkeiten b_i werden nun allerdings als invariant über alle Zeitpunkte angesehen

und als feste Facette mit einer Normierungsbedingung ($\sum_i b_i = 0$) versehen. Für Modell (4.11) wird die durch Item Sampling verursachte Variabilität wiederum mittels Jackknife untersucht. Zusätzlich werden die Items in einem alternativen Modell zu (4.11) als Random Items angenommen, in dem anstelle der bis auf Summennormierung unrestringierten Schätzung der einzelnen Itemschwierigkeiten (anstelle einer nichtinformativen Priorverteilung) eine gemeinsame Normalverteilung der Itemschwierigkeiten angenommen wird. Dieses Modell zieht die einzelnen Itemschwierigkeiten (vor allem bei kleinen Stichproben) tendenziell zur Mitte (siehe die Diskussion zu *No Pooling* und *Partial Pooling* in Gelman & Hill, 2007).

Für die Untersuchung von Item Parameter Drift zwischen den drei Messzeitpunkten werden im mehrdimensionalen Rasch-Modell (4.11) zusätzlich noch zeitpunktspezifische Itemschwierigkeiten b_{it} (mit $\sum_i b_{it} = 0$ für alle $t = 1, 2, 3$) angenommen. Für alle verwendeten IRT-Modelle wäre es auch möglich, zufällige Effekte für Schulklassen und deren Interaktion mit Items und Messzeitpunkten zu modellieren. Empirisch fallen diese Varianzkomponenten im Hinblick auf ihre praktische Bedeutsamkeit jedoch relativ klein aus, so dass auf eine Einführung dieser Parameter in diesem Beitrag verzichtet wird.

Außerdem wird ein mehrdimensionales Rasch-Modell im Rahmen eines *Generalized Linear Mixed Effects Models* spezifiziert, in dem Items und Personen mittels zufälliger Effekte in jeweils einer dreidimensionalen Normalverteilung modelliert werden (sog. *Random Item Profiles*; De Boeck, 2008):

$$\text{logit } \{P(X_{pti} = 1)\} = \mu_t + \theta_{pt} - b_{it} \tag{4.12}$$

Die Restriktion bezüglich der Itemschwierigkeiten bit in Modell (4.12) ergibt sich dadurch, dass für jedes t der Erwartungswert der zufälligen Effekte b_{it} gleich Null ist. Die gerade eingeführten IRT-Modelle wurden mit der Methode Markov Chain Monte Carlo (MCMC) in der frei verfügbaren Software WinBUGS (Spiegelhalter et al., 2003) geschätzt. Die MCMC-Schätzung besitzt den Vorteil, für Effektgrößen exakte statistische Inferenz betreiben zu können und im Gegensatz zu Maximum-Likelihood-Methoden weniger verzerrte Varianzkomponentenschätzungen in der betrachteten Klasse von IRT-Varianzkomponentenmodellen zu ermöglichen (Draper, 2007). Als Parameterschätzer werden für alle Modelle Mittelwerte der eindimensionalen Posteriorverteilungen, als Standardfehlerschätzungen die Standardabweichungen dieser Verteilungen verwendet (Gelman et al., 2004).

Für die imputierten Daten werden zusätzlich separate eindimensionale Skalierungen zu allen Zeitpunkten in der Software ConQuest (Wu, Adams, Wilson & Haldane, 2007) vorgenommen. Während die mehrdimensionalen Modelle eine modellimplizite Verlinkung der Lesekompetenzen in einer konkurrenten Kalibrierung ermöglicht, müssen die zeitpunktspezifischen Skalen der separaten Kalibrierung mit einer anschließenden Verlinkungsmethode auf eine gemeinsame Metrik transformiert werden (Kolen & Brennan, 2004). Dabei gehen in einem linearen Regressionsmodell die Itemparameter der drei Messzeitpunkte in einer Verallgemeinerung des Mean-Mean-Equatings mit Items und Messzeitpunkten als feste Faktoren ein (Haberman, 2009). Diese Verlinkungsmethode kann als Restriktion von Parametern in einer Maximum-Likelihood-Schätzung unter Restriktion von Mittelwerten der Itemschwierigkeiten zu allen drei Zeitpunkten angesehen werden (von Davier & von Davier, 2007). Für die Durchführung einer adäquaten statistischen Inferenz wird die

separate Skalierung mit anschließendem Linking für jeden imputierten Datensatz unter Jackknife jeder Schulklasse (als Cluster) durchgeführt, so dass sich für jeden Datensatz ein Standardfehler ergibt. Auf Basis dieser imputierten Datensätze erfolgt die Inferenz aufgrund der fehlenden Daten nach den Kombinationsregeln von Rubin (Lüdtke & Robitzsch, 2010).

Für die Datenaufbereitung, Bestimmung der Effektgrößen und die Berechnung der statistischen Inferenz mit Resampling-Methoden wurde die Software R (R Core Team, 2014) eingesetzt.

4.6 Ergebnisse

4.6.1 Deskriptive Befunde: Effektgrößen und Stabilitäten

Die auf den Rohwerten basierenden Effektgrößen und Stabilitäten für die drei Messzeitpunkte sind in Tabelle 5.2 dargestellt. Für die Erfassung der Veränderung von T1 nach T2 (d. h. vom Ende der 3. Klasse bis zur Mitte der 4. Klasse) fallen die parametrischen Effektgrößen relativ ähnlich aus ($d = .638$, $d^* = .658$). Die auf Rangdaten operierenden Effektgrößen fallen etwas höher aus ($V^* = .796$, $V^{**} = .675$). Die Differenzen zwischen diesen Effektgrößen weisen auf Abweichungen von der Normalverteilung der Rohwerte hin. Für die längsschnittliche Veränderung von T1 nach T3 (der Entwicklung vom Ende der 3. Klasse bis zum Ende der 4. Klasse) variieren die Effektgrößen zwischen .950 und 1.178. Diese Unterschiede sind größer als die durch Person Sampling bedingte Variabilität statistischer Schätzer.

Tabelle 4.1: *Deskriptive Statistiken und Effektgrößen für den ELFE-Test*

	Parameter	Schätzung	SE_{Pers}	$SE_{Pers-Cl}$	SE_{Items}
$T1 \rightarrow T2$	d	.638	.018	.020	.048
	d^*	.658	.020	.022	.037
	V^*	.796	.020	.021	.023
	V^{**}	.675	.019	.021	.028
$T1 \rightarrow T3$	d	.950	.024	.027	.085
	d^*	1.006	.029	.032	.043
	V^*	1.178	.026	.030	.022
	V^{**}	1.044	.026	.029	.022
	Cor(T1,T2)	.747	.011	.012	.028
	Cor(T1,T3)	.650	.015	.017	.025
	Cor(T2,T3)	.719	.015	.017	.034

Anmerkungen: SE_{Pers}: Standardfehlerschätzung für Person Sampling unter Annahme der unabhängigen Ziehung von Schülern; $SE_{Pers-Cl}$: Standardfehlerschätzung für Person Sampling unter Berücksichtigung der Clusterstruktur (Stichprobenziehung von Klassen); SE_{Items}: Standardfehlerschätzung für Item Sampling.

Die auf Bootstrap basierenden Standardfehler (SE_{Pers}) sind etwas kleiner als die mittels Jackknife einzelner Klassen ermittelten Standardfehler ($SE_{Pers-Cl}$). Im Mittel fallen

die die Clusterstruktur berücksichtigenden Standardfehler etwa 10% höher aus. Diese Faustregel kann für die Adjustierung von Standardfehlern der im Folgenden berichteten IRT-Modelle verwendet werden, die von unabhängigem Sampling von Personen ausgehen. Die relativ geringe Bedeutung der Clusterstruktur der Daten wird durch niedrige Intraklassenkorrelationen (ICC) bestätigt (ICC(T1) = .085, ICC(T2) = .050, ICC(T3) = .069). Aufgrund dieser Befunde wird daher in den folgenden Modellen auf eine explizite Behandlung der Clusterstruktur im Rahmen von Mehrebenenmodellen oder eine implizite Behandlung durch Adjustierung der Standardfehler verzichtet. Tabelle 5.2 entnimmt man, dass die durch Item Sampling ermittelten Standardfehler für die Effektgrößen in den meisten Fällen größer als die durch Person Sampling resultierenden sind. Beispielsweise ist die Effektgröße $d(T1 \rightarrow T3) = .950$ mit einem durch Item Sampling bedingten Standardfehler von $SE_{Items} = .085$ versehen. Die verschiedenen Items messen daher die längsschnittliche Veränderung in unterschiedlichem Ausmaß. Es ist außerdem erkennbar, dass die auf der gepoolten Standardabweichung basierende Effektgröße d^* geringer durch die konkrete Auswahl von Items beeinflusst wird ($SE_{Items} = .043$) als d. Auch für die Stabilitäten zwischen den einzelnen Zeitpunkten sind die durch Item Sampling bedingten Standardfehler durchweg größer als die durch Person Sampling bedingten. Beispielsweise ist der Standardfehler für die Korrelation zwischen T2 und T3 (Cor(T2,T3)=.719) durch Item Sampling ($SE_{Items} = .034$) doppelt so groß wie der durch Person Sampling bedingte ($SE_{Pers-Cl} = .017$).

4.6.2 Varianzkomponentenmodelle

In den Varianzkomponentenmodellen der Generalisierbarkeitstheorie werden zwei Modelle mit logistischer und identischer Linkfunktion spezifiziert und die Sensitivität der Effektgrößen untersucht. In Tabelle 4.2 sind die Schätzungen der Standardabweichungen der zufälligen Facetten und der Effektgrößen sowie deren Standardfehler dargestellt. Die Effektgrößen der Veränderung werden im logistischen Modell etwas geringer als im Modell mit identischer Linkfunktion geschätzt ($d = .642$ im Vergleich zu $d = .710$ für $T1 \rightarrow T2$). Die generellen Aussagen über Varianzanteile der einzelnen Facetten sind jedoch in beiden Modellen relativ ähnlich ausgeprägt. Die Interaktionen von Item und Zeit (d.h. Item Parameter Drift) sowie Schüler und Item (d.h. die überzufällige Tendenz einer gleichen Itemantwort durch einen Schüler zu den Zeitpunkten auf demselben Item) fallen im Modell mit identischer Linkfunktion höher als mit der logistischen Linkfunktion aus (Item x Zeit: 1.4% vs. 4.7% bzw. Schüler x Item: 7.7% vs. 17.3%). Zusammenfassend leitet sich aus diesen Modellen ab, dass von einer Variation in Mittelwertveränderungen in Scores durch Item Sampling ausgegangen werden kann.

4.6.3 Item Parameter Drift

Um das Ausmaß des Item Parameter Drifts (IPD) festzustellen, wurde ein dreidimensionales Rasch-Modell mit über die Messzeitpunkte variierenden Itemschwierigkeiten spezifiziert. Der Mittelwert aller Itemschwierigkeiten wurde dabei als Normierungsbedingung separat zu T1, T2 und T3 gleich Null gesetzt. Aus drei Itemschwierigkeiten eines Items zu den drei Zeitpunkten lässt sich die relative Itemschwierigkeit eines Items zu einem

Tabelle 4.2: *Varianzkomponentenmodelle für den ELFE-Test*

Standard-abweichung		Logistische Linkfunktion			Identische Linkfunktion		
		Est.	SE_{Pers}	Varianz (in %)	Est.	SE_{Pers}	Varianz (in %)
Item	σ_i	2.117	.374	46.8	.208	.037	41.8
Item x Zeit	σ_{it}	.373	.048	1.4	.070	.008	4.7
Schüler	σ_p	1.836	.036	35.2	.171	.003	28.3
Schüler x Zeit	σ_{pt}	.858	.020	8.9	.091	.002	7.9
Schüler x Item	σ_{pi}	.922	.025	7.7	.134	.002	17.3
Residuum	σ_{pti}	1.814	.000[§]	—	.315	.001	—

Anmerkungen: [§] In Modellen mit der logistische Linkfunktion ist die Standardabweichung des Residuums durch die Standardabweichung der logistischen Verteilung gegeben. Est.: Parameterschätzung

Zeitpunkt als Differenz von Itemschwierigkeit zu diesem Zeitpunkt und der mittleren Itemschwierigkeit dieses Items über alle drei Zeitpunkte bestimmen.

In Abbildung 4.2 sind für jeden der drei Messzeitpunkte die relativen Itemschwierigkeiten in Abhängigkeit von der Itemposition im ELFE-Test dargestellt. Negative relative Itemschwierigkeiten bedeuten dabei, dass entsprechende Items zum jeweiligen Messzeitpunkt für die Schüler leichter sind, positive relative Itemschwierigkeiten, dass diese Items zum jeweiligen Messzeitpunkt schwieriger sind.

Die einfacheren Items an den vorderen Positionen des Tests fallen dabei zu T1 tendenziell einfacher, zu T3 schwieriger aus. Dies wird durch die Korrelation der mittleren Itemschwierigkeiten (Mittelwert der Itemschwierigkeiten zu T1, T2 und T3) mit den relativen Itemschwierigkeiten der einzelnen Zeitpunkte bestätigt ($r(T1) = .86$, $r(T2) = -.46$, $r(T3) = -.81$). Der Leistungsunterschied zwischen T1 und T3 würde demzufolge unter Verwendung der Items an den Positionen 1 bis 10 geringer, bei Verwendung der Items der Positionen 11 bis 20 größer ausfallen. Das Ausmaß des Item Parameter Drifts lässt sich als Standardabweichung aller Differenzen relativer Itemschwierigkeiten verschiedener Zeitpunkte ausdrücken. Dabei ergeben sich als Standardabweichungen für IPD SD(T1→T2) = .35, SD(T1→T3) = .55 und SD(T2→T3) = .26. Je weiter die Messzeitpunkte auseinander liegen, desto größer ist daher die Unsicherheit der ermittelten Effektgrößen aufgrund differenziellen Itemfunktionierens zu einzelnen Zeitpunkten. Dieser Befund zeigt, dass in Abhängigkeit von den ausgewählten Items mittlere Veränderungen heterogen ausfallen und daher von einem Einfluss des Item Sampling auf Effektgrößen auszugehen ist.

4.6.4 Effektgrößen aus mehrdimensionalen Rasch-Modellen

Die in Tabelle 5.2 und 4.2 berichteten Befunde sollen nun mit denen aus einem mehrdimensionalen Rasch-Modell mit zeitinvarianten Itemschwierigkeiten verglichen werden. In Modell M1 werden die Itemschwierigkeiten als invariant über die Zeitpunkte und als Fixed Items, in Modell M2 als invariant und als Random Items angenommen. Die Invarianzannahme wird in den Modellen M3 und M4 fallen gelassen: die Annahme Fixed Items wird

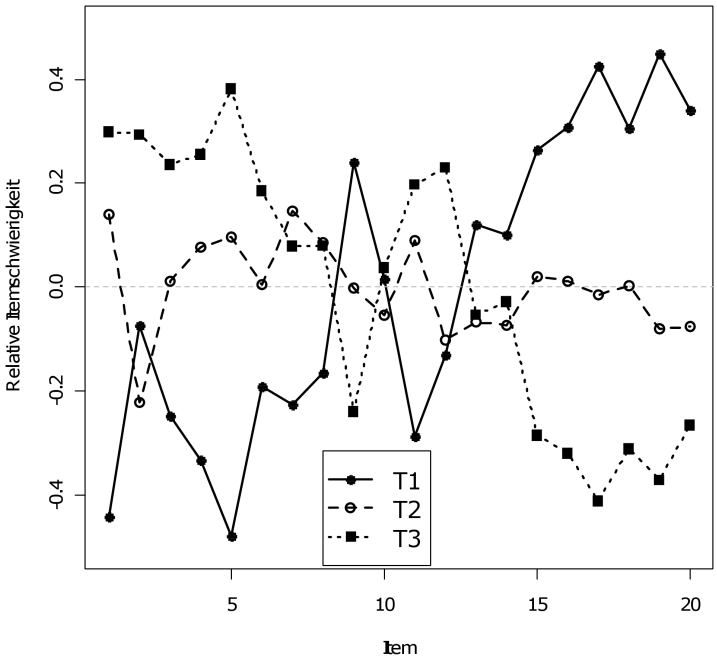

Abbildung 4.2: *Relative Itemschwierigkeiten der 20 Items im ELFE-Test zu den Zeitpunkten T1 (Ende 3. Klasse), T2 (Mitte 4. Klasse) und T3 (Ende 4. Klasse).*

in Modell M3 durch separate Kalibrierungen mit anschließendem Linking vorgenommen, Random Items in Modell M4 werden mit einer dreidimensionalen Normalverteilung für Itemschwierigkeiten operationalisiert.

In Tabelle 4.3 erkennt man zunächst für Modell M1, dass sich die Effektgrößen d und d^* insbesondere für die Messung der Veränderung von T1 nach T3 deutlich unterscheiden ($d = 1.324$, $d^* = 1.131$). Die Wahl der zu T1 und T3 gepoolten Standardabweichung für die Berechnung der Effektgröße d^* führt aufgrund der zu T3 deutlich größeren Standardabweichung im Vergleich zu T1 zu einer praktischen Verringerung des Effekts. In geringerem Ausmaß wird dieses Phänomenen auch für T1→T2 sichtbar. Für beide Veränderungen stellt man aber fest, dass die durch Item Sampling bedingten Standardfehler bei d größer als für d^* sind (für T1→T3: $SE(d) = .089$ bzw. $SE(d^*) = .032$) . Dieses Ergebnis steht im Einklang mit den Befunden der auf den Rohwerten basierenden Effektgrößen (siehe Tabelle 5.2).

Die Korrelationen zwischen den drei Messzeitpunkten sind mit Cor(T1,T2) = .870, Cor(T1,T3) = .791 und Cor(T2,T3) = .843 in diesem Modell mit latenten Variablen erwartungsgemäß deutlich höher als im Modell mit manifesten Variablen (siehe Tabelle 5.2). Vergleicht man M1 (Fixed Items) mit M2 (Random Items), so stellt man im Hinblick auf Parameterschätzungen nahezu keine Unterschiede bei minimal erhöhten Standardfehlern fest. In den beiden Modellen M3 und M4 bei Nichtinvarianz der Itemschwierigkeiten werden die Effektgrößen geringer geschätzt ($d(T1 \rightarrow T3)$ in M1: 1.324, in M3: 1.071). Hinsichtlich der Parameterschätzungen sind die Modelle M3 und M4 sehr ähnlich. Zusätzlich erkennt man, dass die Standardabweichung der Itemschwierigkeiten SD_{Items} von Zeitpunkt T1 nach T3 in den Modellen M3 und M4 deutlich absinkt. Dieses Phänomen

könnte eine Begründung darstellen, weshalb die Effektgrößen in M3 und M4 geringer als diejenigen in M1 und M2 unter Annahme invarianter Itemschwierigkeiten ausfallen. In Modell M4 fallen die Standardfehler im Vergleich zu allen anderen Modellen am höchsten aus, da Items und Personen immer als zufällig spezifiziert wurden. Da Modell M3 mit imputierten Datensätzen gerechnet wurde, kann der Varianzanteil der Parameterschätzung als fraction of missing information (*fmi*; siehe Lüdtke & Robitzsch, 2010) berechnet werden, der auf fehlende Daten zurückzuführen ist. Für die Effektgrößen d und d^* bewegt sich die *fmi* von .064 bis .103.

Für den Fall der Effektgröße $d(T1 \rightarrow T3)$ soll das Prinzip des Model Averaging an den vier Modellen in Tabelle 4.3 illustriert werden. Ein Forscher sieht dabei die Modelle M1 bis M4 als theoretisch gleich plausibel an, so dass Modellgewichte $w_m = .25 = 1/4$ gewählt werden. Sollen diese Priorwahrscheinlichkeiten der Modelle nicht anhand der empirischen Daten revidiert werden, so werden diese Gewichte im Model Averaging verwendet. Die Effektstärke d aus den vier Modellen ergibt sich dann als Mittelwert aller Modellparameterschätzungen: $d = .25 \cdot (1.324 + 1.325 + 1.071 + 1.073) = 1.198$. Die Varianz der Schätzung d im Rahmen der Multi Model Inference ergibt sich als

$$Var(d)_{Modelle} = .25 \cdot \left[(1.324 - 1.198)^2 + (1.325 - 1.198)^2 + (1.071 - 1.198)^2 + (1.073 - 1.198)^2\right] = .0159 \tag{4.13}$$

so dass der Standardfehler der Effektgröße infolge der Model Uncertainty $\sqrt{.0159} = .126$ beträgt. Diese Größe übersteigt die Standardfehler innerhalb der Modelle, die aufgrund von Person Sampling oder Item Sampling ermittelt werden.

4.7 Diskussion

In diesem Beitrag wurde die längsschnittliche Entwicklung der Lesekompetenz bei Grundschülern mit dem ELFE-Test untersucht. Die konkrete Auswahl von Items (hier als Item Sampling bezeichnet) stellt sich dabei im Hinblick auf Effektgrößen als eine bedeutende Variabilitätsquelle für statistische Parameter dar. Die vorgelegten Befunde zeigen, dass Item Sampling auch unter der Annahme eines sehr großen Stichprobenumfangs von Personen Berücksichtigung erfahren sollte. Wenn eine Generalisierung der Befunde über das konkrete eingesetzte Testmaterial hinaus von Relevanz ist, wovon in der Forschungsliteratur in der Regel auszugehen ist, sollte das heterogene Funktionieren von Items im Hinblick auf Veränderungsmaße durch Standardfehler aufgrund von Item Sampling quantifiziert werden. Empirisch wurde gezeigt, dass die durch die Itemauswahl verursachten Standardfehler mindestens so groß sind wie die durch die Ziehung von Personen verursachten Standardfehler. Aus diesem Blickwinkel scheint die Wahl von Effektgrößen der Veränderung sinnvoll, die einen kleinen Standardfehler aufgrund der Itemauswahl besitzen und daher möglichst insensitiv gegenüber der konkreten Auswahl von Items sind. Rein deskriptiv zeigt sich im Einklang mit den Befunden von Briggs und Weeks (2009), dass die Effektgröße d^*, die an der mittleren Standardabweichung der beiden Messzeitpunkte gepoolt ist, weniger sensitiv gegenüber Item Sampling als die Effektgröße d ist, die eine Standardisierung an der Standardabweichung zu T1 vornimmt.

Für die Messung der mittleren Veränderung von T1 nach T3 (also eines Schuljahres)

Tabelle 4.3: *Parameterschätzungen aus mehrdimensionalen Rasch-Modellen*

	M1: Fixed Items (invariant)				M2: Random Items (invariant)		M3: Fixed Items (nicht invariant)					M4: Random Items (nicht invariant)	
Parameter	Est.	SE_{Pers}	SE_{Item}	SE_{Total}	Est.	SE_{Pers}	Est.	$SE_{Pers-Cl}$	SE_{Miss}	SE_{Total}	*fmi*	Est.	SE
$d(T1 \rightarrow T2)$	.787	.024	.045	.051	.788	.025	.673	.029	.008	.030	.064	.670	.104
$d^*(T1 \rightarrow T2)$	.724	.020	.024	.031	.725	.021	.664	.025	.007	.026	.071	.650	.054
$d(T1 \rightarrow T3)$	1.324	.035	.089	.096	1.325	.037	1.071	.041	.013	.043	.097	1.073	.117
$d^*(T1 \rightarrow T3)$	1.131	.026	.032	.041	1.131	.026	1.031	.036	.012	.037	.103	1.001	.058
$SD(T1)$	1.588	.030	.118	.122	1.586	.031	1.716	.039	.008	.040	.043	1.719	.034
$SD(T2)$	1.852	.037	.209	.212	1.851	.037	1.763	.044	.009	.045	.041	1.820	.036
$SD(T3)$	2.094	.045	.250	.254	2.096	.046	1.846	.043	.015	.046	.106	1.969	.044
$SD_{Items}(T1)$	1.825	.015	.458	.458	1.822	.015	2.053	.043	.008	.044	.032	2.047	.023
$SD_{Items}(T2)$	1.825	.015	.458	.458	1.822	.015	1.746	.043	.014	.045	.091	1.749	.025
$SD_{Items}(T3)$	1.825	.015	.458	.458	1.822	.015	1.591	.049	.017	.052	.114	1.579	.027
Cor(T1,T2)	.870	.009	.021	.023	.869	.009	—	—	—	—	—	.867	.009
Cor(T1,T3)	.791	.012	.027	.030	.792	.013	—	—	—	—	—	.792	.012
Cor(T2,T3)	.843	.011	.025	.027	.844	.011	—	—	—	—	—	.850	.010

Anmerkungen: Est.: Paramterschätzung; SD: Standardabweichung der Lesekompetenzen der Schüler; SD_{Items}: empirische Standardabweichung der geschätzten Itemschwierigkeiten; Cor: Korrelationen der Lesekompetenz zwischen den Zeitpunkten; SE_{Pers}: Standardfehler aufgrund Person Sampling; SE_{Item}: Standardfehler aufgrund Item Sampling; $SE_{Pers-Cl}$: Standardfehler aufgrund Person Sampling unter Berücksichtigung der Clusterstruktur (Jackknife); SE_{Miss}: Standardfehler aufgrund fehlender Daten (Variabilität zwischen multipel imputierten Datensätzen); SE_{Total}: Standardfehler aufgrund Person Sampling und Item Sampling (Modell M1) bzw. aufgrund Person Sampling und fehlenden Daten (Modell M3); *fmi*: Fraction of Missing Information.

ergaben sich in den in dieser Arbeit verwendeten verschiedenen Skalierungsmodellen und Effektgrößen Werte zwischen .950 und 1.324. Diese Spannweite der Maße übersteigt um ein Vielfaches die aufgrund der Ziehung von Personenstichproben erwartungsgemäße Variabilität von Effektgrößen, die mit Hilfe von Standardfehlern berechnet wird. Das hier berichtete Intervall $[.950, 1.324]$ kann daher als Bereich plausibler Effektgrößen aufgefasst werden, in dem sich auch schon in der Literatur berichtete Befunde des verwendeten ELFE-Tests befinden ($d = 1.07$; Pfost, Karing, Lorenz & Artelt, 2010).

Wenn Item Sampling Variabilität in Effektgrößen generiert, kann Item Parameter Drift eine Ursache darstellen. Die Items messen somit nicht homogen längsschnittliche Veränderung. Im vorliegenden ELFE-Test könnten die tendenziell zum Ende des Tests vorgelegten Items aufgrund von Testermüdungseffekten für die Drittklässler (relativ) schwieriger als für die Viertklässler ausfallen.

Die Überlappung der Verteilung von Itemschwierigkeiten mit der Verteilung von Personenfähigkeiten führt dabei in Abhängigkeit von der Wahl einer Linkfunktion in einem IRT-Modell (z.B. die Verwendung einer logistischen oder einer identischen Linkfunktion) zu einer Änderung der Standardabweichung zu T2 oder T3, die damit zu einem gewissen Ausmaß willkürlich eine Effektgröße beeinflussen kann. Die Prüfung von Annahmen über die Verteilung einer latenten Variablen im IRT-Modelle – wie der Konstanz der Varianz der Schülerfähigkeiten über die Zeitpunkte hinweg – hängt von der Verteilung der Itemschwierigkeiten und der Wahl der Linkfunktion ab, so dass diese Tests immer nur für ein konkretes Modell durchgeführt werden können. Wenn a priori kein Modell aus theoretischen Gründen zu präferieren ist, halten wir diese Art von Modelltestungen als fragwürdig. Umgekehrt könnte beispielsweise ein Forscher aber empirisch ein IRT-Modell ermitteln, welches näherungsweise konstante Varianzen der Schülerkompetenzen generiert, da der Forscher postuliert, dass diese IRT-Scores eine höhere Validität als Scores aus alternativen Modellen besitzen (Kane, 2006).

Nichtparametrische deskriptive Maße der Veränderung wie V^* setzen nur die Ordinalskaliertheit von Scores voraus und sind daher weniger stark von der Wahl einer (IRT-)Metrik oder konkreter Itemmengen (d.h. Item Sampling) betroffen. Die in diesem Artikel berichteten nichtparametrischen Maße sollten daher in Studien den parametrischen Maßen (wie d oder d^*) gegenübergestellt werden. Auf manifesten Rohwerten basierende Effektgrößen wie V^* geben im Beispiel des ELFE-Tests sogar recht gute Abschätzungen für zu erwartende Effektgrößen aus IRT-Modellen mit latenten Variablen. Daher sollte im Rahmen einer Sensitivitätsanalyse eine Inferenz über verschiedene theoretisch plausibler Modelle vorgenommen werden, um Aussagen über interessierende Effektgrößen zu erhalten. Wenn latente Variablen in statistischen Modellen auftreten, argumentieren wir, dass die Wahl eines (Skalierungs-)Modells, die aus substanziellen oder aus Gründen maximaler statistischer Effizienz basierenden Erwägungen erfolgt, nicht zwangsläufig mit der Modellwahl übereinstimmt, die die Maximierung des globalen Modellfits als Zielstellung besitzt. Die Untersuchung des Modellfits wird in vielen Fällen unter Annahme von Fixed Items und einer Random Person Perspektive vorgenommen. Die verschiedenen Komponenten der Verteilung von Itemschwierigkeiten, Personenfähigkeiten und der Wahl der Linkfunktion sind aber für die Beurteilung eines Modellfits nicht unabhängig voneinander. Maris und Bechger (2009) zeigen beispielsweise, dass ohne Spezifikation der Verteilung von Personenfähigkeiten in der Fixed Persons Perspektive ein Rasch-Modell empirisch nicht von einem

IRT-Modell mit Rateparametern unterscheidbar ist. Für Längsschnittanalysen heißt dies konsequenterweise, dass ein Forscher definieren muss, ob in der Definition von Kompetenzen mit Hilfe latenter Variablen Rateverhalten in einem IRT-Modell modelliert werden soll oder nicht. Ein Modellselektionskriterium leistet in dieser Fragestellung daher nur bedingt Hilfestellung. Die in diesem Beitrag berichteten Modelle M1 bis M4 unterscheiden sich darin, ob Itemschwierigkeiten als zeitinvariant oder Items als zufällig angenommen werden. Aufgrund von Item Parameter Drift und starken Unterschieden der Streuung von Itemschwierigkeiten in den Modellen M3 und M4 könnte man als Forscher mit Hilfe eines Modellfitkriteriums M3 und M4 gegenüber M1 oder M2 präferieren. In einem solchen Modellvergleich würde man aber in allen Modellen annehmen, dass das IRT-Modell für alle Personen gültig ist und die Random Persons Perspektive zutrifft. Anstelle eines differenziellen Funktionierens von Items könnte man analog auch von differenziellem Funktionieren von Personen (*Person Misfit*; Meijer & Sijtsma, 2001) ausgehen und somit sogar noch komplexere Modelle zulassen. Es ist hier stark zu betonen, dass eine bessere Passung eines alternativen komplexen Modells mit einem besseren globalen Modellfit auch die Bedeutung des interessierenden Parameters (z. B. Effektgrößen der Veränderung) verändern kann (etwa in einem Modells, worin die lokalen Abhängigkeiten von zu einem Stimulus zugehörigen Items mit Hilfe eines Testletfaktors modelliert werden). Der Forscher sollte dann ein Modell wählen, mit dessen Hilfe er für diesen Parameter Aussagen bezüglich angestrebter theoretischer Annahmen treffen kann. Im Hinblick auf Item Sampling wurde in diesem Beitrag argumentiert, dass bei einer gewünschten Generalisierung der konkret eingesetzten Items auf eine größere Population von Items die Annahme von festen Items nicht notwendigerweise getroffen werden sollte. Der aus statistischen Gründen vorgenommene Ausschluss nichtmodellkonformer Items könnte deshalb eine Einschränkung der Repräsentativität nach sich ziehen, weil keine „repräsentative Stichprobenziehung von Items“ mehr erfolgt.

Mit der Wahl einer logistischen Linkfunktion gegenüber der identischen Linkfunktion sollen Lösungswahrscheinlichkeiten in einem extremen Wahrscheinlichkeitsbereich stärker als im mittleren Wahrscheinlichkeitsbereich gespreizt werden (Goldstein, 1980). Ob man Veränderung durch diese Art der Operationalisierung erfassen will, entscheidet der Forscher, denn in IRT-Modellen können nicht Item-Response-Funktionen einer beliebig monotonen Form nichtparametrisch und unabhängig von der Spezifikation einer Fähigkeitsverteilung der latenten Variablen geschätzt werden (Douglas, 2001; vgl. auch Alonso, Litière & Laenen, 2010). Erst wenn die Form der Fähigkeitsverteilung fixiert wird (etwa eine Normalverteilung oder eine Gleichverteilung), können komplexere Item-Response-Modelle als das Rasch-Modell identifiziert werden (siehe hierzu Haberman, 2005). In vielen Anwendungen scheint allerdings zumindest die monotone Transformierbarkeit von IRT-Scores gewährleistet (d.h. die Ordinalskaleneigenschaft), so dass die auf diesen Annahmen basierenden Effektgrößen mit möglichst wenig Verteilungs- und Modellannahmen zu präferieren sind. Sowohl empirisch als auch konzeptuell zeigen diese Überlegungen, dass die aus der so genannten spezifischen Objektivität des Rasch-Modells (bei dessen Gültigkeit) folgende „Unabhängigkeit von Itemparametern“ bezüglich der Personenstichprobe und die „Unabhängigkeit von Personenparametern“ bezüglich der Itemstichprobe fragwürdig ist (siehe hierzu für eine typische Darstellung aus der Rasch-Perspektive Bond & Fox 2001 sowie für eine kritische Diskussion dieses Ansatzes van der Linden, 1994, 2001). Insbesondere

die Annahme der Unabhängigkeit der Itemauswahl setzt für die Ermittlung individueller Fähigkeitsschätzungen voraus, dass das verwendete IRT-Modell für jeden Testteilnehmer gültig ist, also die Wahrscheinlichkeiten $P(X_{pi} = 1)$ für jede einzelne Person p und jedes Item i interpretiert werden kann und daher diese Person als ein *Stochastic Subject* operiert (vgl. hierzu die kontrastierende Perspektive des *Random Sampling* in Holland, 1990a; Wainer, 2010b). Praktisch schließt man dabei aus dem Vorliegen interindividueller Querschnittsdaten (Responses von N Personen zu einem Zeitpunkt für allen Items) auf die intraindividueller Gültigkeit des IRT-Modells für alle Personen (siehe Molenaar, 2004 oder Alisch, 2002 für eine Diskussion). Jede Form itemspezifischer Varianz wird dann als intraindividuelle Fehlerquelle interpretiert. Zumindest für Kompetenztests ist diese Annahme kritisch zu hinterfragen, da man bei wiederholtem Vorlegen eines Tests für Schüler bei bestimmten Items eher deterministisches als probabilistisches Antwortverhalten erwarten kann (Holland, 1990a; Molenaar, 1995; Wainer, 2010b). Für statistische Analysen mit Modellselektion wurde im Beitrag argumentiert, dass „konventionelle statistische Inferenz" für das ausgewählte Modell im Allgemeinen Standardfehler für interessierende Parameter unterschätzt. Die integrierte Betrachtung im Rahmen einer Multi Model Inference mit a priori gewählten theoretisch plausiblen Modellen könnte eine geeignete Alternative darstellen.

Aus Vereinfachungsgründen der Darstellung wurde die Varianzzerlegung im Konzept der Generalisierbarkeit nur mit Haupteffekten unter Aussparung von Interaktionen vorgenommen. Werden Testteilnehmern dieselben Items in Längsschnittstudien vorgelegt, so ist mit einer substanziellen Varianzquelle der Interaktion von Person und Item aufgrund lokaler Abhängigkeiten auszugehen. In diesem Fall muss die Forschungsfragestellung entscheiden, ob diese Abhängigkeiten expliziter Modellbestandteil sein sollen (Marais, 2009) oder ob diese lokale Abhängigkeiten nicht im IRT-Modell Berücksichtigung finden, aber Standardfehler von Modellparametern adjustiert werden müssen (Ip, 2010).

Neben Modellwahl und Itemwahl kann im vorliegenden Datensatz auch die Modellierung des Drop-Out von Schülern eine Unsicherheitsquelle darstellen. Die verwendeten IRT-Modelle beruhen entweder auf modellbasierten Verfahren aus beobachteten Daten oder auf imputierten Datensätzen unter der Annahme Missing at Random (MAR). Bestehen mehrere Forschungsfragestellungen für denselben Datensatz, so ist nach unserer Meinung eine multiple Imputation dem modellbasierten Vorgehen vorzuziehen, da Annahmen über den Ausfallprozess separat von interessierenden Analysemodellen spezifiziert und für die folgenden Analysen fixiert werden. Analysen auf Basis imputierter Datensätzen ermöglichen im Gegensatz zum modellbasierten Vorgehen eine explizite Abschätzung der Varianz von Parameterschätzungen aufgrund fehlender Daten (Yucel, 2008). Abweichungen von der Annahme Missing at Random können im R-Paket mice im Sinne einer Sensitivitätsanalyse der Abweichung von MAR umgesetzt werden (van Buuren & Groothuis-Oudshoorn, 2011; Resseguier, Roch & Paoletti, 2011).

Für Datensätze mit sehr großen Personenstichprobenumfängen wie in Large-Scale Assessments (z.B. der PISA-Studie) können im Vergleich zum Person Sampling verschiedene andere systematische und unsystematische Fehlerquellen bedeutsamer sein (Wu, 2010; Wuttke, 2007). Diese Behauptung findet im Beispiel der Lesekompetenzentwicklung anhand des ELFE-Tests Bestätigung.

Die angestellten Überlegungen könnten weitreichende Implikationen für wichtige Fra-

gestellungen der Pädagogischen Psychologie und der Entwicklungspsychologie besitzen. Ein häufig diskutiertes Problem ist die Existenz des so genannten Schereneffektes in der Sekundarstufe (Becker, Lüdtke, Trautwein & Baumert, 2006; Pfost et al., 2010; Retelsdorf & Möller, 2008). Hierbei wird davon ausgegangen, dass die nach dem Übergang in die Sekundarstufe bestehende Differenz der Leistungen der Schüler in den unterschiedlichen Schulformen im Verlauf der Zeit weiter zunimmt. Die in den o.g. Referenzen verwendeten Effektgrößen basieren auf parametrischen Effektstärkenmaßen und setzen (mindestens implizit) Intervallskalenniveau der verwendeten Scores voraus. Da diese Studien eine Verankerung der Standardabweichung zu T1 aufgrund der Verwendung von d vornehmen, könnte es sein, dass sich in der Literatur berichtete Schereneffekte mit anderen Skalierungsmodellen, anderen Effektstärkenmaßen und anderem Itemmaterial als nicht stabil erweisen und daher zum Teil eine Konsequenz von Item- und Modellwahl sind. Möglichst voraussetzungsarme statistische Modelle (wobei explizit auch Modelle mit manifesten Variablen einzuschließen sind) können dabei einen guten Eindruck über die zu erwartende Sensitivität von Effektgrößen geben. Neben dem durch Person Sampling verursachten Standardfehler wurde anhand der Lesekompetenzentwicklung mit dem ELFE-Test gezeigt, dass im Konzept der Generalisierbarkeit die beiden Facetten Items und verwendete Modelle interessierende statistische Parameter in einem nicht vernachlässigbaren Ausmaß beeinflussen können. Die Publikation dieser Variation oder deren „Standardfehler“ hilft, Generalisierbarkeit von Effektgrößen über den Rahmen einer Einzelstudie hinaus zu ermöglichen.

Kapitel 5

Zur (Nicht-)Modellierung lokaler Abhängigkeiten in Messmodellen

5.1 Einleitung

In diagnostischen Tests liegen häufig mehrere Items zu einem gemeinsamen Stimulus vor (*Testlet*) oder sind anderweitig zu verschiedenen Gruppen (z.B. Subskalen) zusammengefasst. In der Literatur wird darauf hingewiesen, dass in Testlets vorliegende Items eine Abhängigkeit generieren, die in eindimensionalen Messmodellen mit einer Verletzung der Annahme der lokalen stochastischen Unabhängigkeit verbunden ist (siehe z.B. Wainer et al., 2007; Tuerlinckx und De Boeck, 2004; Eckes, 2015a, 2014). Lokale stochastische Unabhängigkeit meint hierbei, dass sich eine Verteilung $P(X_1, \ldots, X_I)$ von I Items als ein Produkt von Verteilungen einzelner Items bedingt auf eine latente Variable θ parametrisieren lässt (Holland, 1990a)

$$P\left[(X_1, \ldots, X_I) = (x_1, \ldots, x_I)\right] = \int_\theta \prod_{i=1}^{I} P_i(x_i|\theta) g(\theta)\, \mathrm{d}\theta \tag{5.1}$$

wenn P_i die Item-Response-Funktion von Item i und g eine Dichtefunktion einer latenten Variable θ bezeichnet[1]. Die Verteilung der latenten Variable θ ist dabei erst durch die Annahme der lokalen stochastischen Unabhängigkeit in einem Messmodell definiert (McDonald, 1981).

Wenn lokale Abhängigkeiten ignoriert werden würden, führt dies nach Ansicht vieler Autoren zu einer Überschätzung der Traitvarianz bzw. der Itemtrennschärfen und damit verbunden zu einer überschätzten Reliabilität des Skalenwertes (DeMars, 2006; Ip, 2000). Um diese Abhängigkeiten angemessen in einem psychometrischen Modell zu berücksichtigen, wurden Testlet-Modelle (Wainer et al., 2007; Eckes, 2015a; Min & He, 2014) oder Bifaktor-Modelle (Reise, 2012; Gignac, 2014; Brunner, Nagy & Wilhelm, 2012) vorgeschlagen, die in den letzten beiden Jahrzehnten größere Verbreitung in der psychometrischen Literatur gefunden haben (siehe für eine Übersicht Rauch & Moosbrugger, 2011). Nach der Logik dieser Ansätze werden verschiedene faktorenanalytische Modelle an die Daten

[1] Die latente Variable kann dabei auch diskret verteilt sein. In diesem Fall geht die Integration in eine Summation über.

angepasst und dann auf Basis des am besten passenden Modells die Reliabilität bestimmt. Die Berechnung der Reliabilität wird somit empirisch über die Anpassung eines Faktormodells begründet.

In dem vorliegenden Beitrag wird argumentiert, dass die Frage nach der Wahl eines psychometrischen Modells zur Bestimmung der „richtigen“ Reliabilität nicht ausschließlich empirisch begründet werden sollte. Aus dieser Sicht würde eine sehr gute Passung des Testlet-Modells auf die Daten sowie eine bedeutsame Varianz der Testletfaktoren noch nicht zwangsläufig zur Konsequenz haben, dass zur Berechnung der Reliabilität ein Testlet-Modell (bzw. Bifaktor-Modell) verwendet werden sollte. Es wird der Standpunkt vertreten, dass die Reliabilität eines Test über die Spezifikation einer hypothetischen Testreplikation zu definieren ist (Brennan, 2001a; Kane, 2011). Für den Anwender stellt sich somit die Frage, über welche Bedingungen (z.B. Situation, Teststimuli etc.) der Testwert generalisiert werden soll. Der Vergleich des Fits verschiedener faktorenanalytischer Modelle ist dagegen für die Beantwortung dieser Frage nur wenig aussagekräftig. Des Weiteren wird betont, dass die Reliabilität eines Tests nicht losgelöst von seiner Validität betrachtet werden sollte.

Zunächst diskutieren wir in Abschnitt 5.2 mit Hilfe eines Zahlenbeispiels die Schätzung der Reliabilität anhand verschiedener Modelle, die mit jeweils unterschiedlichen Annahmen die Abhängigkeiten zwischen den Items berücksichtigen. In Abschnitt 5.3 werden die verschiedenen psychometrische Modelle anhand eines realen Datenbeispiels (HAMLET-Test) illustriert. In Abschnitt 5.4 erweitert die Perspektive und zieht die Validität eines Tests in die Überlegungen mit ein. Abschließend wird in Abschnitt 5.5 ein kurzer Ausblick auf alternative Modellierungen gegeben sowie diskutiert, welche Konsequenzen sich aus unseren Überlegungen für die Evaluation psychometrischer Modelle ergeben.

5.2 Psychometrische Modellierung lokaler Abhängigkeiten

5.2.1 Drei Ansätze zur Modellierung lokaler Abhängigkeit

Im Folgenden sollen die Überlegungen an einem fiktiven Beispieldatensatz illustriert werden. Es wird von einem Test mit insgesamt sechs Items (A1, A2, B1, B2, C1, C2) ausgegangen, wobei jeweils zwei Items einem Testlet zugeordnet sind, so dass insgesamt drei Testlets A, B und C resultieren. Es bieten sich drei alternative Modelle zur Analyse der Testlets an (siehe Abbildung 5.1). In Modell M1 wird eine eindimensionale Fähigkeit θ durch die Annahme der lokalen stochastischen Unabhängigkeit definiert. In Modell M2 werden zusätzliche Testletfaktoren u_A, u_B und u_C aufgenommen, die die Abhängigkeit der Items in den Testlets modellieren. In Modell M3 werden die Abhängigkeiten innerhalb von Testlets in einem eindimensionalen Modell mit korrelierten Residuen abgebildet.

Es sei nun eine (tetrachorische) Korrelationsmatrix $\mathbf{\Sigma}$ vorgegeben, für die angenommen wird, dass die Items verschiedener Testlets weniger stark zusammenhängen ($r = .40$) als

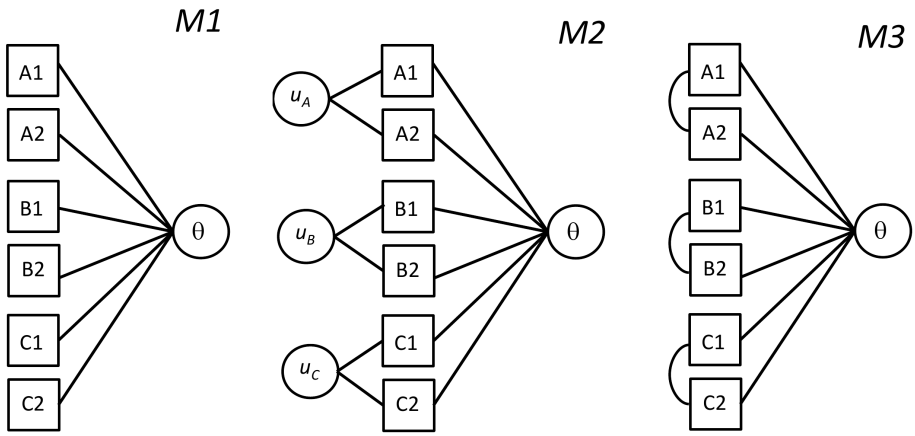

Abbildung 5.1: *Verschiedene Messmodelle zur Modellierung von Abhängigkeiten zwischen Items. Links: Eindimensionales Modell mit unkorrelierten Fehlern (M1); Mitte: Testlet-Modell (bzw. Bifaktor-Modell; M2); Rechts: Eindimensionales Modell mit positiven lokalen Abhängigkeiten (M3)*

die Items innerhalb von Testlets ($r_{A_1A_2} = .55$, $r_{B_1B_2} = .60$, $r_{C_1C_2} = .50$):

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & .55 & .4 & .4 & .4 & .4 \\ .55 & 1 & .4 & .4 & .4 & .4 \\ .4 & .4 & 1 & .6 & .4 & .4 \\ .4 & .4 & .6 & 1 & .4 & .4 \\ .4 & .4 & .4 & .4 & 1 & .5 \\ .4 & .4 & .4 & .4 & .5 & 1 \end{pmatrix} \tag{5.2}$$

Die leicht höheren Korrelationen für Items innerhalb eines Testlets sind charakteristisch für Tests mit einer Testlet-Struktur. Für die folgenden Überlegungen soll davon ausgegangen werden, dass diese Korrelationsstruktur in der Population gilt[2]. Die durch eine Stichprobenziehung entstehende Unsicherheit soll nicht weiter berücksichtigt werden.

Die Modelle M1, M2 und M3 werden mit der Unweighted Least Squares (ULS) Schätzmethode (Savalei, 2014; ten Berge, 1993) unter der Annahme konstanter Itemtrennschärfen angepasst. In der ULS-Schätzmethode wird die Summe der quadrierten Differenz aus beobachteten Kovarianzen σ_{ij} und erwarteten Kovarianzen $\hat{\sigma}_{ij}$ betrachtet, d.h. die Funktion $F = \sum_{i \neq j}(\sigma_{ij} - \hat{\sigma}_{ij})^2$ minimiert. Dabei wird für jedes der Modelle M1, M2 und M3 die ermittelte Traitvarianz sowie die mit Hilfe der ω-Maße (Reise et al., 2010; Zinbarg, Revelle, Yovel & Li, 2005) geschätzte Reliabilität berichtet. Modell M1 geht von gleichen Trennschärfen und unkorrelierten Residuen aus (Modell tau-äquivalenter Messungen). Es wird folgende Modellanpassung vorgenommen: $\mathbf{\Sigma} \approx \mathbf{\Phi} + \mathbf{\Theta}$ Dabei bezeichnet $\mathbf{\Phi}$ die Kovarianzmatrix der Faktorstruktur (der Varianz der wahren Werte) und $\mathbf{\Theta}$ die residuale Kovarianzmatrix. Die ULS-Schätzmethode liefert dann folgende Zerlegung der Korrelationsmatrix

[2]Die Betrachtung einer Korrelationsmatrix ist nicht als Einschränkung zu sehen, da bei dichotomen Daten die Verwendung des Probitlinks (der Normalverteilungsfunktion) äquivalent zur Anpassung eines Faktormodells an eine tetrachorische Korrelationsmatrix ist (Kamata & Bauer, 2008)

$$\boldsymbol{\Sigma} \approx \boldsymbol{\Phi} + \boldsymbol{\Theta} = \begin{pmatrix} .43 & .43 & .43 & .43 & .43 & .43 \\ .43 & .43 & .43 & .43 & .43 & .43 \\ .43 & .43 & .43 & .43 & .43 & .43 \\ .43 & .43 & .43 & .43 & .43 & .43 \\ .43 & .43 & .43 & .43 & .43 & .43 \\ .43 & .43 & .43 & .43 & .43 & .43 \end{pmatrix} + \begin{pmatrix} .57 & 0 & 0 & 0 & 0 & 0 \\ 0 & .57 & 0 & 0 & 0 & 0 \\ 0 & 0 & .57 & 0 & 0 & 0 \\ 0 & 0 & 0 & .57 & 0 & 0 \\ 0 & 0 & 0 & 0 & .57 & 0 \\ 0 & 0 & 0 & 0 & 0 & .57 \end{pmatrix} \tag{5.3}$$

Die geschätzte Traitvarianz beträgt $\phi = .43$ und die Kovarianzmatrix der wahren Werte lässt sich als $\boldsymbol{\Phi} = \phi \mathbf{1}\mathbf{1}'$ schreiben, wobei $\mathbf{1}$ ein aus lauter Einsen bestehenden Vektor bezeichnet. Für den Summenscore wird die *klassische Reliabilität* ρ nach Lucke (2005) definiert als

$$\rho = \frac{\mathbf{1}'\boldsymbol{\Phi}\mathbf{1}}{\mathbf{1}'\boldsymbol{\Phi}\mathbf{1} + \mathbf{1}'\boldsymbol{\Theta}\mathbf{1}} \tag{5.4}$$

Die Reliabilität ergibt sich somit als der Anteil der Varianz der wahren Werte an der Gesamtvarianz. Im Fall eines kongenerischen (und damit auch eines tau-äquivalenten) Modells erhält ρ die Bezeichnung ω (McDonald, 1999) [3]. Dabei ergibt sich in Modell M1 eine Reliabilitätsschätzung für den Summenscore von $\omega = .82$, die mit Cronbachs Alpha übereinstimmt. Das Modell M1 weist jedoch einen bedeutsamen Misfit auf. Zur Bestimmung des Ausmaßes der Fehlspezifikation (sog. *Modellfehler*; Cudeck & Henly, 1991; MacCallum, 2003) wird der Approximationsfehler $\boldsymbol{\Delta} = \boldsymbol{\Sigma} - (\boldsymbol{\Phi} + \boldsymbol{\Theta})$ berechnet

$$\boldsymbol{\Delta} = \begin{pmatrix} 0 & .12 & -.03 & -.03 & -.03 & -.03 \\ .12 & 0 & -.03 & -.03 & -.03 & -.03 \\ -.03 & -.03 & 0 & .17 & -.03 & -.03 \\ -.03 & -.03 & .17 & 0 & -.03 & -.03 \\ -.03 & -.03 & -.03 & -.03 & 0 & .07 \\ -.03 & -.03 & -.03 & -.03 & .07 & 0 \end{pmatrix} \tag{5.5}$$

Der Modellfehler wird durch die Fitstatistik $SRMR = \sqrt{\left(\sum_{i \neq j} \Delta_{ij}^2\right) / (I \cdot (I-1))}$ quantifiziert. Für Modell M1 ergibt sich $SRMR = .06$.Es sei allerdings bereits auf eine Eigenschaft der geschätzten Residualkorrelation Δ_{ij} hingewiesen, diespäter noch eine zentrale Rolle spielen wird. In der ULS-Schätzmethode wird die quadrierte Summe der Modellfehler minimiert, d.h. $F = \sum_{i \neq j} (\sigma_{ij} - \hat{\sigma}_{ij})^2 = \sum_{i \neq j} \Delta_{ij}^2$. Analog zur linearen Regression ist die Summe dieser Residuen Δ_{ij} gleich Null (Savalei, 2014). In diesem Modell heben sich daher die geschätzten positiven lokalen Abhängigkeiten (zwischen Items innerhalb eines Testlets) und negativen lokalen Abhängigkeiten (zwischen Items verschiedener Testlets) auf: Summe $2 \cdot (.12 + .17 + .07) - 24 \cdot (-.03) = 0$ (siehe Habing & Roussos, 2003 für ähnliche Befunde in IRT-Modellen).

Das Modell M2, in dem zusätzlich drei Testletfaktoren spezifiert werden, besitzt einen perfekten Modellfit (SRMR = .00). Es ergibt sich eine Traitvarianz von .40 sowie spezifische Testletvarianzen von .15, .20 und .10. Als Modellanpassung erhält man

[3] Für die wesentliche Argumentation in diesem Beitrag ist es nicht bedeutsam, ob im datengenerierenden Modell von gleichen Itemladungen ausgegangen wird. Die Hauptaussagen des Beitrags bleiben auch mit verschiedenen Itemladungen bestehen.

$$\boldsymbol{\Sigma} = \boldsymbol{\Phi} + \boldsymbol{\Theta} = \begin{pmatrix} .55 & .55 & .4 & .4 & .4 & .4 \\ .55 & .55 & .4 & .4 & .4 & .4 \\ .4 & .4 & .6 & .6 & .4 & .4 \\ .4 & .4 & .6 & .6 & .4 & .4 \\ .4 & .4 & .4 & .4 & .5 & .5 \\ .4 & .4 & .4 & .4 & .5 & .5 \end{pmatrix} + \begin{pmatrix} .45 & 0 & 0 & 0 & 0 & 0 \\ 0 & .45 & 0 & 0 & 0 & 0 \\ 0 & 0 & .4 & 0 & 0 & 0 \\ 0 & 0 & 0 & .4 & 0 & 0 \\ 0 & 0 & 0 & 0 & .5 & 0 \\ 0 & 0 & 0 & 0 & 0 & .5 \end{pmatrix} \tag{5.6}$$

In der Matrix $\boldsymbol{\Phi}$, die Varianzen und Kovarianzen der vier Faktoren abbildet, befindet sich in der Diagonalen jeweils die Summe aus Traitvarianz und Testletvarianz (.55, .6 und .5). Für die Berechnung der Reliabilität müssen zwei Ansätze unterschieden werden.

Zum einen kann sowohl die Varianz des generellen Faktors θ als auch die Testletvarianz als Bestandteil der Varianz des wahren Wertes (Omega Total, ω_t) angesehen werden. Dies würde bedeuten, dass alle vier Faktoren in unserem Beispiel wahre Varianz erfassen. Zum anderen kann die Testletvarianz als Bestandteil der Fehlervarianz (Omega Hierarchical, ω_h) betrachtet werden (Reise et al., 2010; Reise, 2012; siehe auch Rauch & Moosbrugger, 2011). In diesem Fall wird nur die Varianz des Traitfaktors (.40) als wahre Varianz interpretiert. Für unser Beispiel ergeben sich deutlich verschiedene Reliabilitäten von $\omega_t = .86$ und $\omega_h = .76$. Es stellt sich die Frage, welcher von den beiden Werten nun als Reliabilität des Tests interpretiert werden soll. Nach Wainer und Thissen (1996) ist die Angabe der Reliabilität ω_t mit der Annahme verbunden, dass die Testletfaktoren konstruktinhärent sind und eine hypothetische Testreplikation zu derselben Testletstruktur führen würde (*feste Testlet-Dimensionalität*). Dagegen geht ω_h von der Annahme aus, dass die Testlets konstruktirrelevant sind (*zufällige Testlet-Dimensionalität*). D.h. bei einer erneuten Durchführung des Tests könnte der Test auch eine andere Testletstruktur aufweisen. Diese Unterscheidung entspricht der Interpretation der Stufen eines Faktors in der Varianzanalyse als „fest“ oder „zufällig“ (Maxwell & Delaney, 2004; Searle, Casella & McCulloch, 1992).

Diese Unterscheidung lässt sich beispielsweise anhand des Tests für Deutsch als Fremdsprache illustrieren. Der TestDaF-Test weist eine Test-Struktur auf, da jeweils zu unterschiedlichen Text-Stimuli mehrere Aufgaben vorgelegt werden. In einer Analyse der psychometrischen Eigenschaften des TesDaF wurden von Eckes (Eckes, 2015a) die verschiedenen Testlets als feste Faktoren interpretiert und die Reliabilität des TestDaF nur auf Basis der Traitvarianz berechnet. Alternativ könnte hier auch argumentiert werden, dass die Testlets konstruktinhärent sind, da die einzelnen Testlet-Stimuli beispielsweise verschiedene Textsorten (Kurztext, Zeitungsbericht, wissenschaftlicher Artikel), kognitive Anforderungen (Kompetenzniveaus) oder Itemformate (Matching, Multiple Choice, True-False-Not Given) repräsentieren, die auch bei einer erneuten Durchführung Bestandteil des Tests wären.

Im letzten Schritt wird das Modell M3 angepasst, in dem die Residuen der Items eines Testlets als korreliert angenommen werden. Das Modell weist ebenso einen perfekten Modellfit ($SRMR = .00$) auf und führt zu einer Traitvarianz von .40 sowie einer Reliabilität von $\omega_h = .76$. Es lässt sich zeigen, dass das Modell M3 statistisch äquivalent zu dem mehrdimensionalen Testlet-Modell M2 ist (siehe Ip, 2010 für IRT-Modelle). Dabei entspricht die Reliabilität dem Modell M2 mit konstruktirrelevanten Testlets ($\omega_h = .76$),

d.h. wenn die Testletvarianz zur Fehlervarianz gerechnet wird. Das Modell M3 führt dabei folgende Kovarianzzerlegung durch

$$\boldsymbol{\Sigma} = \boldsymbol{\Phi} + \boldsymbol{\Theta} = \begin{pmatrix} .4 & .4 & .4 & .4 & .4 & .4 \\ .4 & .4 & .4 & .4 & .4 & .4 \\ .4 & .4 & .4 & .4 & .4 & .4 \\ .4 & .4 & .4 & .4 & .4 & .4 \\ .4 & .4 & .4 & .4 & .4 & .4 \\ .4 & .4 & .4 & .4 & .4 & .4 \end{pmatrix} + \begin{pmatrix} .6 & .15 & 0 & 0 & 0 & 0 \\ .15 & .6 & 0 & 0 & 0 & 0 \\ 0 & 0 & .6 & .2 & 0 & 0 \\ 0 & 0 & .2 & .6 & 0 & 0 \\ 0 & 0 & 0 & 0 & .6 & .1 \\ 0 & 0 & 0 & 0 & .1 & .6 \end{pmatrix} \tag{5.7}$$

Dieses Modell besitzt demzufolge mittlere positive lokale Abhängigkeiten, da die Summe der Nichtdiagonalelemente von $\boldsymbol{\Theta}$ positiv ist (denn $2 \cdot .15 + 2 \cdot .2 + 2 \cdot .1 = .90$).

Insgesamt lässt sich somit die Frage nach der „richtigen“ Reliabilität nicht empirisch entscheiden, sondern verlangt vom Anwender die Definition einer hypothetischen Testreplikation (Brennan, 2001c). Mit dieser Definition wird gleichzeitig festgelegt, ob die Testletdimensionen als fest oder zufällig angesehen werden. Keinesfalls sollten bedeutsame Testletvarianzen im Testlet-Modell (M2) dahingehend interpretiert werden dürfen, dass nur das Testlet-Modell bzw. ω_h eine unverzerrte Schätzung der Reliabilität liefert und bei der Wahl anderer Modelle (z.B. dem eindimensionalen IRT-Modell) eine verzerrte Reliabilität resultieren würde. Des Weiteren gilt es zu berücksichtigen, dass eine Unterschätzung der Reliabilität – z.B. durch eine nicht angemessene Verwendung von ω_h – keineswegs immer als konservativ angesehen werden kann. Setzt man beispielsweise eine unterschätzte Reliabilität für die Berechnung einer latenten (messfehlerbereinigten) Korrelation zu einer Kovariaten ein, so würde sich eine Überschätzung dieser Korrelation ergeben.

5.2.2 Domain Sampling

Der Ansatz des *Domain Samplings* (Cronbach, 1951; Cronbach & Shavelson, 2004; Nunnally & Bernstein, 1994, S. 211ff.; Brennan, 2001a) stellt eine alternative Perspektive auf die Bestimmung der Reliabilität dar. Wie in diesem Abschnitt deutlich werden wird, lässt sich aus dieser Perspektive auch die Wahl des weniger gut passenden Modells M1 zur Bestimmung der Reliabilität rechtfertigen. Es lässt sich zeigen, dass Modell M1, das die lokalen Abhängigkeiten ignoriert, einer Schätzung der Reliabilität mittels Cronbachs Alpha entspricht. Für eine gegebene Kovarianzmatrix $\boldsymbol{\Sigma}$ ist die *interne Konsistenz* nach Cronbachs Alpha gegeben als (Lucke, 2005)

$$\alpha = \frac{I}{I-1} \cdot \left(1 - \frac{\mathrm{tr}\boldsymbol{\Sigma}}{\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}}\right) = \frac{I^2\phi}{\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}} \tag{5.8}$$

wobei tr den Trace-Operator bezeichnet, der als Summe der Diagonalelemente einer Matrix definiert ist und ϕ den Mittelwert der Nichtdiagnonalelemente in $\boldsymbol{\Sigma}$ bezeichnet.

In der Literatur zu Cronbachs Alpha (z.B. Raykov & Marcoulides, 2010; Steyer & Eid, 2001; Gignac, 2014) dominiert die Sicht, dass zur Bestimmung der Reliabilität die Annahmen der Eindimensionalität, gleicher Itemtrennschärfen und unkorrelierter Residuen erfüllt sein müssen (Modell tau-äquivalenter Messungen). Für den *Domain Sampling* Ansatz (Cronbach, 1951; Cronbach & Shavelson, 2004; Nunnally & Bernstein, 1994, S. 211ff.; Brennan, 2001a) stellen diese Annahmen keine Voraussetzung für die Bestimmung

der Reliabilität mittels Cronbachs Alpha dar. In diesem Ansatz wird lediglich die Annahme getroffen, dass in einer hypothetischen Testreplikation (derselben definierten Domäne) Items dieselbe mittlere Kovarianz aufweisen (Tryon, 1957). D.h. es werden keine expliziten Annahmen über die Dimensionalität der Items getroffen.

Die Idee des Domain Samplings und einer nichtdiagonalen Residualkovarianzmatrix $\boldsymbol{\Theta}$ ist in Abbildung 5.2 dargestellt.

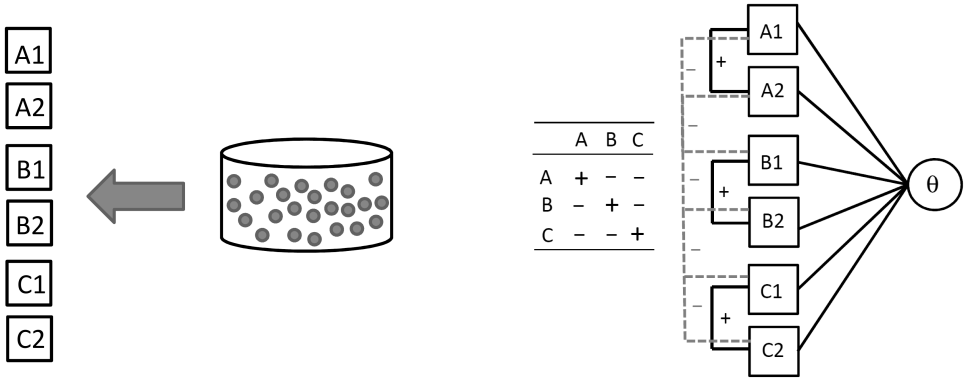

Abbildung 5.2: *Links: Item Sampling aus einer Domain; Rechts: Eindimensionales Messmodell mit positiv und negativ korrelierten Residuen*

Die Idee einer (ggf. hypothetischen) Ziehung von Items aus einer Itempopulation wird in verschiedenen Teilgebieten der psychometrischen Forschung diskutiert, wie z.B. der Generalisierbarkeitstheorie (Cronbach, Gleser, Nanda & Rajaratnam, 1972; Brennan, 2001a), Arbeiten zur psychometrischen Inferenz (Husek & Sirotnik, 1967; Hunter, 1968; Mulaik, 2009a), dem Behavior Domain Sampling (McDonald, 2003; Markus & Borsboom, 2013a) oder in theoretischen Überlegungen zur Existenz latenter Variablen in IRT-Modellen (Stout, 1990; Junker, 1991, 1993). Insbesondere für Längsschnittstudien wird dabei Item Sampling als Variabilitätsquelle für die Messung von Veränderung angesehen (Michaelides, 2010; Michaelides & Haertel, 2004; Strathmann & Klauer, 2010; Robitzsch et al., 2011).

Im Rahmen des Domain Sampling Ansatzes wäre ein Modellfit eines faktorenanalytischen Modells nicht von Bedeutung, da lediglich die Annahme eines unabhängigen (hypothetischen) Item-Samplings zu treffen ist (Brennan, 2001a). Das bedeutet, dass in einem Messmodell (nichtmodellierte) positive und negative lokale stochastische Abhängigkeiten existieren können, sich aber ausmitteln und das eindimensionale Modell M1 unter Ignorierung der lokalen Abhängigkeiten mit nichtperfektem Fit zur „richtigen“ Reliabilität führen kann. In der rechten Grafik in Abbildung 5.2 wird schematisch dargestellt, dass bei Anpassung eines Einfaktormodells Residualkorrelationen von Items innerhalb eines Testlets positiv und von Items verschiedener Testlets negativ ausfallen. Wir werden auf die Bedeutung dieser Grafik noch einmal im Rahmen der approximate factor models in Abschnitt 5.2.3 zurückkommen.

Die Beobachtung, dass sich positive und negative Residualkorrelationen im Ansatz des Domain Samplings ausmitteln, lässt sich auch formal mit Hilfe der Definition der klassischen Reliabilität ρ zeigen. Bezeichnet ϕ wiederum die mittlere Residualkovarianz und

wird die Kovarianzmatrixzerlegung $\mathbf{\Sigma} = \mathbf{\Phi} + \mathbf{\Theta}$ mit $\mathbf{\Phi} = \phi\mathbf{1}\mathbf{1}'$ und $\mathbf{\Theta} = \mathbf{\Sigma} - \mathbf{\Phi}$ verwendet, so folgt $\alpha = \rho$, d.h. die Reliabilität wird durch Cronbachs Alpha erwartungstreu geschätzt:

$$\rho = \frac{\mathbf{1}'\mathbf{\Phi}\mathbf{1}}{\mathbf{1}'\mathbf{\Phi}\mathbf{1} + \mathbf{1}'\mathbf{\Theta}\mathbf{1}} = \frac{\mathbf{1}'(\phi\mathbf{1}\mathbf{1}')\mathbf{1}}{\mathbf{1}'(\phi\mathbf{1}\mathbf{1}')\mathbf{1} + \mathbf{1}'\mathbf{\Theta}\mathbf{1}} = \frac{I^2\phi}{\mathbf{1}'\mathbf{\Sigma}\mathbf{1}} = \alpha \tag{5.9}$$

Hier gilt es zu beachten, dass nicht die Frage nach der Passung eines Modells gestellt wird, sondern eine Modellpassung per Definition gegeben ist. Im Fall unkorrelierter Fehler kann man zeigen, dass die Reliabilität $\rho = \omega$ immer mindestens so groß wie Cronbachs Alpha ausfällt und daher α eine untere Schranke für ω darstellt (Lucke, 2005; Revelle & Zinbarg, 2009). Bei korrelierten Fehlern kann α sowohl größer als auch kleiner als ω ausfallen. Die Beziehung hängt von der Unterschiedlichkeit der Ladungen sowie den Korrelationen der Fehler ab (siehe Lucke, 2005).

Im Ansatz des Domain-Samplings ist es nicht von Bedeutung, ob ein Modell kongenerischer Messungen (mit verschiedenen Itemladungen) besser als ein Modell tau-äquivalenter Messungen passt. Die zu erfassende latente Variable θ wird als gleichgewichtete Repräsentation der Items *definiert*. Dies steht im Kontrast zu einer fitbasierten und empirisch extrahierten Ableitung der Reliabilität, wenn sich Itemladungen (und damit „Beiträge“ der Items für die latente Variable) unterscheiden.

Formal kann ein Domain Sampling Modell durch folgende Modellgleichung dargestellt werden (siehe Brennan, 2001a)

$$X_{pi} = \mu + \theta_p - b_i + e_{pi} \tag{5.10}$$

wobei Items mit i und Personen mit p bezeichnet werden [4]. Die statistische Inferenz wird dabei im Hinblick auf eine Population von Personen *und* Items betrieben. Aussagen über „Personenparameter“ θ_p beziehen sich dabei auf eine Itempopulation und über „Itemparameter“ b_i auf eine Personenpopulation. Die Frage, ob die Spezifikation von unterschiedlichen Itemtrennschärfen a_i in (5.10) „notwendig“ ist, stellt sich hier nicht, da diese zusätzlichen Parameter nicht als Bestandteil des datengenerierenden Modells im Domain Sampling angesehen werden. Würde man diese Itemtrennschärfen etwa mit der Nebenbedingung eines Mittelwertes von 1 (d.h. $E(a_i) = 1$) in (5.10) einfügen, ergäbe sich folgende Modellgleichung

$$X_{pi} = \mu + a_i(\theta_p - b_i) + e_{pi} = \mu + \theta_p - b_i + \underbrace{(a_i - 1)(\theta_p - b_i) + e_{pi}}_{=:\tilde{e}_{pi}} \tag{5.11}$$

wobei $\tilde{e}_{pi}$ eine neue Fehlervariable definiert. Demzufolge sind in (5.10) „verschieden funktionierende“ Items Bestandteil der Interaktionsvariablen e_{pi} von Personen und Items. In

[4] Ohne Einschränkung der Allgemeinheit könnte man in (5.10) auch die in IRT-Modellen oft verwendete logistische Linkfunktion für die Behandlung dichotomer oder ordinaler Daten einfügen (Glas, 2012b; Briggs & Wilson, 2007). Die Verwendung der Modellgleichung (5.10) für dichotome Items in Verbindung mit der Schätzung der kleinsten Quadrate bedeutet *nicht*, dass Modellparameter im linearen Modell im Gegensatz zum logistischen Modell „verzerrt“ sind (Greene, 2003). Die Wahl einer konkreten Linkfunktion definiert vielmehr die Metrik der interessierenden Variablen θ (Ballou, 2009; Robitzsch et al., 2011). D.h. wenn Unterschiede zwischen Personen in der Originalmetrik (der Summenscore-Metrik) abgebildet werden sollen, so kommt (trotz dichotomer Items!) die identische Linkfunktion zum Einsatz.

Analogie zu Items könnten in (5.10) auch Personentrennschärfen a_p (Repräsentation von Person Misfit; vgl. Meijer & Sijtsma, 2001) eingefügt werden (z.B. Raiche et al., 2013; Agarwal, Zhang & Mazumder, 2011; Khanna, Zhang, Agarwal & Chen, 2013). Im Domain Sampling würde Heterogenität in diesen Parametern aber wiederum in der Fehlervariablen e_{pi} subsumiert werden. Aus der Perspektive des Domain Samplings ist es deshalb nicht ganz verständlich, warum dem Misfit von Items in Faktormodellen oder IRT-Modellen so eine prominente Stellung eingeräumt wird, während der Passung des Modells für einzelne Personen nur relativ wenig Aufmerksamkeit gewidmet wird. In diesem Sinne wählt Brennan (2011) für den Vergleich von klassische Testtheorie (CTT) bzw. Generalisierbarkeitstheorie und Item-Response-Theorie (IRT) folgende Metapher:

> A forest-trees metaphor is reasonably apt for considering IRT vis-a-vis expected value theories. Consider *individual items* as *trees* and the *universe of items* as the *forest*. If we focus on individual trees as we do in IRT, then we are easily oblivious to the forest. If we focus on the forest [CTT], then the trees are *indistinguishable*.
>
> To put it another way, in IRT *items* (more correctly item parameters) are effectively *fixed*, which means that a replication would consist of identically the same items (or, more correctly, a set of items with identically the same parameters). Call this „strictly" parallel forms. The notion of randomly parallel forms in [CTT] is much less restrictive, and even the various CTT notions of parallel forms are much weaker than „strictly" parallel forms.

Items werden im Domain Sampling (bzw. der CTT) nach Brennan (2011) als nicht unterscheidbar (bzw. austauschbar) angesehen, in der IRT jedoch nicht. Für die Testkonstruktion hat dies die Implikation, dass bei der Untersuchung von Testeigenschaften nur auf den *forest* und nicht die einzelnen *trees* Augenmerk gelegt wird. Im Ansatz des Domain Samplings ist man somit an einer Aussage über die Menge aller Items und aller Personen im Hinblick auf eine Population interessiert.

5.2.3 Rolle des Modellfehlers bei der Schätzung der Reliabilität

Im Folgenden wollen wir die Rolle des Modellfehlers $\boldsymbol{\Delta}$ bei der faktorenanalytischen Schätzung der Reliabilität etwas näher beleuchten. Sowohl in der Psychometrie (Cudeck & Henly, 1991; MacCallum, Browne & Cai, 2007) als auch der Ökonometrie (z.B. Chamberlain & Rothschild, 1983; Bai & Ng, 2002; Bai & Ng, 2008; Bai & Li, 2012a; siehe auch Press & Shigemasu, 1997; Rowe, 2002, Kap. 9) existieren Arbeiten zu Faktormodellen mit Modellfehlern $\boldsymbol{\Delta}$, die sog. *Approximate Factor Models*, in denen davon ausgegangen wird, dass Modelle „nur approximativ" in der Population gelten. So treffen MacCallum und Kollegen (2007) die Annahmen, dass neben den dominanten Faktoren eine Vielzahl kleiner Faktoren („weak factors") bestehen, die zu einem Modellfehler führen und an denen man nicht weiter inhaltlich interessiert ist.

Um die Bedeutung des Modellfehlers zu illustrieren, schreiben wir zunächst die Formel zur Definition der Reliabilität ρ um

$$\rho = \frac{\mathbf{1}'\boldsymbol{\Phi}\mathbf{1}}{\mathbf{1}'\boldsymbol{\Phi}\mathbf{1} + \mathbf{1}'\boldsymbol{\Theta}\mathbf{1}} = \frac{\mathbf{1}'\boldsymbol{\Phi}\mathbf{1}}{\mathbf{1}'\boldsymbol{\Phi}\mathbf{1} + \mathrm{tr}\boldsymbol{\Theta} + \mathrm{tr}(\boldsymbol{H}\boldsymbol{\Theta})} \tag{5.12}$$

Dabei ist $\boldsymbol{H} = \mathbf{1}\mathbf{1}' - \boldsymbol{I}$ eine Matrix, die in den Nichtdiagonalelementen Einsen und in den Diagonalelementen Nullen enthält. Der Ausdruck $\mathrm{tr}(\boldsymbol{H}\boldsymbol{\Theta})$ ist dabei die Summe der residualen Kovarianzen. Wenn sich positive und negative lokale Abhängigkeiten ausmitteln, so folgt $\mathrm{tr}(\boldsymbol{H}\boldsymbol{\Theta}) = 0$ und man erhält aus (5.12)

$$\rho = \frac{\mathbf{1}'\boldsymbol{\Phi}\mathbf{1}}{\mathbf{1}'\boldsymbol{\Phi}\mathbf{1} + \mathrm{tr}\boldsymbol{\Theta}} \tag{5.13}$$

Theoretisch wird demzufolge von einer unbekannten Kovarianzzerlegung $\boldsymbol{\Sigma} = \boldsymbol{\Phi} + \boldsymbol{\Theta}$ ausgegangen. Wird nun ein konkretes Modell angepasst (z.B. ein IRT-Modell an einen Test mit Testletstruktur) trifft man vereinfachende Annahmen, so dass man eine Zerlegung $\boldsymbol{\Sigma} = \tilde{\boldsymbol{\Phi}} + \tilde{\boldsymbol{\Theta}} + \boldsymbol{\Delta}$ mit einem Modellfehler $\boldsymbol{\Delta}$ erhält. Die Matrizen $\tilde{\boldsymbol{\Phi}}$ und $\tilde{\boldsymbol{\Theta}}$ sind die für die Population bestimmten Modellparameter. Wird das Modell mit der ULS-Schätzmethode angepasst, so folgt, dass die mittlere lokale Abhängigkeit $\mathbf{1}'\boldsymbol{\Delta}\mathbf{1}$ gleich Null ist. D.h. aufgrund der Modellanpassung wird implizit die Identifikationsannahme getroffen, dass sich positive und negative lokale Abhängigkeiten ausmitteln. Diese Annahme ist deutlich schwächer als die Forderung, dass alle paarweisen Residualkovarianzen gleich Null sind (lokale stochastische Unabhängigkeit).

Führt man diese Überlegungen weiter, so ergibt sich eine im Faktormodell geschätzte Reliabilität $\tilde{\rho}$ von

$$\tilde{\rho} = \frac{\mathbf{1}'\tilde{\boldsymbol{\Phi}}\mathbf{1}}{\mathbf{1}'\tilde{\boldsymbol{\Phi}}\mathbf{1} + \mathrm{tr}\tilde{\boldsymbol{\Theta}} + \mathrm{tr}(\boldsymbol{H}\tilde{\boldsymbol{\Theta}})} \tag{5.14}$$

Es können dann hinreichende Bedingungen abgeleitet werden, wann die geschätzte Reliabilität $\tilde{\rho}$ in (5.14) mit der unbekannten (theoretischen) Reliabilität ρ in (5.12) übereinstimmt. Zwei Bedingungen, die zusammen hinreichend sind, wären gegeben durch

(B1) $\mathbf{1}'\boldsymbol{\Phi}\mathbf{1} = \mathbf{1}'\tilde{\boldsymbol{\Phi}}\mathbf{1}$, also $\mathbf{1}'(\boldsymbol{\Phi} - \tilde{\boldsymbol{\Phi}})\mathbf{1} = 0$.

(B2) $\mathbf{1}'\boldsymbol{\Delta}\mathbf{1} = 0$

Die Bedingung (B2) ist eine Identifikationsannahme für die Schätzung des Modells und bedeutet, dass sich Modellfehler ausmitteln sollen. Die Bedingung (B1) bedeutet, dass die geschätzte wahre Varianz $\mathbf{1}'\tilde{\boldsymbol{\Phi}}\mathbf{1}$ sich im Mittel nicht von der theoretischen wahren Varianz $\mathbf{1}'\boldsymbol{\Phi}\mathbf{1}$ unterscheidet.

Aus den Bedingungen (B1) und (B2) kann man wegen $\boldsymbol{\Sigma} = \boldsymbol{\Phi} + \boldsymbol{\Theta} = \tilde{\boldsymbol{\Phi}} + \tilde{\boldsymbol{\Theta}} + \boldsymbol{\Delta}$ die Beziehung (F3) folgern

(F3) $\mathrm{tr}\boldsymbol{\Theta} + \mathrm{tr}(\boldsymbol{H}\boldsymbol{\Theta}) = \mathrm{tr}\tilde{\boldsymbol{\Theta}} + \mathrm{tr}(\boldsymbol{H}\tilde{\boldsymbol{\Theta}})$

Das bedeutet, dass die mittlere geschätzte Residualkovarianz gleich der mittleren theoretischen Residualkovarianz ist. Typischerweise wird $\tilde{\boldsymbol{\Theta}}$ in Modellanpassungen eine relativ einfache Struktur besitzen und daher als Schätzer für $\boldsymbol{\Theta}$ fehlspezifiziert sein. Betrachten wir die in Anwendungen häufig getroffenen Annahme unkorrelierter Residuen, so ist $\mathrm{tr}(\boldsymbol{H}\tilde{\boldsymbol{\Theta}}) = 0$. Da die Diagonalen der angepassten Residualkovarianzmatrizen wegen (B1) im Allgemeinen übereinstimmen werden (und damit $\mathrm{tr}\boldsymbol{\Theta} = \mathrm{tr}\tilde{\boldsymbol{\Theta}}$), folgt daraus $\mathrm{tr}(\boldsymbol{H}\boldsymbol{\Theta}) = 0 = \mathbf{1}'\boldsymbol{\Delta}\mathbf{1}$. D.h. die nichtmodellierte wahre Residualkovarianzmatrix $\boldsymbol{H}\boldsymbol{\Theta}$ wird durch den Modellfehler so abgebildet, dass der Mittelwert der nichtmodellierten wahren

Residualkovarianz gleich dem Mittelwert der Modellfehler ist. Das korrespondierende Faktormodell des Domain Sampling, das zur Reliabilität nach Cronbachs Alpha führt, passt die Faktorkovarianzmatrix $\tilde{\mathbf{\Phi}} = \phi \mathbf{1}\mathbf{1}'$ und eine diagonale Residualkovarianzmatrix $\tilde{\mathbf{\Theta}}$ an eine Kovarianzmatrix $\mathbf{\Sigma}$ mit der ULS-Schätzmethode an[5]. Als Identifikationsbedingung für den Modellfehler $\mathbf{\Delta}$ wird dann $\mathbf{1}'\mathbf{\Delta}\mathbf{1} = 0$ angenommen.

Zusammenfassend kann festgehalten werden, dass sich Modellfehler und lokale Abhängigkeiten nicht empirisch voneinander trennen lassen. Zur Identifikation des faktorenanalytischen Modells müssen Annahmen über beide getroffen werden. Die hier vorgeschlagene Interpretation zielt darauf ab, mit einer Modellanpassung die „primäre Dimension“ zu identifizieren, die den Daten zugrunde liegt (siehe auch Ip, Molenberghs, Chen, Goegebeur & De Boeck, 2013). Aus praktischer Sicht stellt lokale stochastische Unabhängigkeit demnach eine nicht realisierbare (und auch nicht in allen Anwendungen sinnvolle) Modellannahme dar.

5.2.4 Unbestimmtheit der Reliabilität und Modelläquivalenz

Während in Abschnitt 5.2.1 und Abschnitt 5.2.3 die Reliabilität mit Hilfe von (approximativen) Faktormodellen gewonnen wurde, beruhte die Bestimmung der Reliabilität in Abschnitt 5.2.2 nur auf einer Matrixzerlegung $\mathbf{\Sigma} = \mathbf{\Phi} + \mathbf{\Theta}$, die ohne Annahmen über ein Faktormodell auskam. Abschließend soll auf eine radikale Position eingegangen werden, die sich aufgrund des Problems äquivalenter Modelle ergibt. In dieser Position wird die Ansicht vertreten, dass die Reliabilität im Rahmen eines Faktormodells nicht eindeutig bestimmt ist und lediglich ein Bereich „plausibler“ Reliabilitäten bestimmt werden kann (Westfall, Henning & Howell, 2012). In Abschnitt 5.2.1 wurde bereits deutlich, dass die Modelle M2 (Testlet-Modell) und M3 (eindimensionales Modell mit korrelierten Residuen) statistisch äquivalent sind, d.h. sie besitzen denselben Modellfit und sind empirisch voneinander nicht abgrenzbar. Die Modelle M2 und M3 besitzen demzufolge Matrixzerlegungen

$$\text{M2} : \mathbf{\Sigma} = \mathbf{\Phi}_2 + \mathbf{\Theta}_2 \quad \text{bzw.} \quad \text{M3} : \mathbf{\Sigma} = \mathbf{\Phi}_3 + \mathbf{\Theta}_3 \tag{5.15}$$

In den Gleichungen (5.15) treten dabei keine Modellfehler auf, da es sich um Modelle mit perfektem Fit handelt. Zur Vereinfachung der nachfolgenden Überlegungen gehen wir

[5]Bei Modellfehlspezifikation kann dabei die ULS-Schätzmethode (ULS) robuster als die Maximum-Likelihood-Schätzung (ML) bei fehlspezifizierten Modellen sein (MacCallum et al., 2007). Während ULS die Zielfunktion $F_{ULS} = \sum_{i,j}(\sigma_{ij} - \hat{\sigma}_{ij})^2$ minimiert, minimiert ML (näherungsweise) $F_{ML} \approx \sum_{i,j}(\sigma_{ij} - \hat{\sigma}_{ij})^2/(u_i^2 u_j^2)$, wobei u_i die Varianz des Itemresiduums ist (MacCallum et al., 2007, S. 166). Für Itempaare (i, j) mit hoher Kommunalität werden daher in der ML-Schätzung die Abweichungen stärker gewichtet. Eine robuste Alternative zur ULS-Schätzmethode stellt die LAD-Schätzmethode dar, die die Summe der betraglichen Abweichungen als Zielfunktion minimiert: $F_{LAD} = \sum_{i,j} |\sigma_{ij} - \hat{\sigma}_{ij}|$ (Siemsen & Bollen, 2007). Es ist zu betonen, dass daher bei Abweichung vom Modell tau-äquivalenter Messungen die Maximum-Likelihood-Schätzung zu anderen Parameterschätzern als die ULS-Schätzmethode und damit auch zu einer anderen Reliabilität führt, da ML und ULS verschiedene Zielfunktionen optimieren (MacCallum et al., 2007).

Spielen Modellfehler bei der Anpassung von IRT-Modellen bei dichotomen oder ordinalen Daten eine Rolle, so sind nach derselben Überlegung sog. Limited-Information-Maximum-Likelihood Schätzmethoden (Pairwise-Marginal-Maximum-Likelihood, siehe Renard et al., 2004 oder McDonald, 1997) Full-Information-Maximum-Likelihood Methoden überlegen (z.B. Marginal-Maximum-Likelihood Schätzung in IRT-Software).

zunächst nicht auf Modellfehler ein.

Westfall et al. (2012) argumentieren, dass ohne konkrete Annahmen an $\boldsymbol{\Phi}$ und $\boldsymbol{\Theta}$ die Anpassung eines Faktormodells und damit verbunden die Schätzung der Reliabilität unbestimmt ist. Um diesen Gedanken zu illustrieren, nehmen wir an, dass eine Matrixzerlegung in einem Modell Ma der Form $\boldsymbol{\Sigma} = \boldsymbol{\Phi}_a + \boldsymbol{\Theta}_a$ gültig sei. Ein äquivalentes Modell Mb ist dadurch konstruierbar, dass man zu $\boldsymbol{\Phi}_a$ eine Matrix $\boldsymbol{U}$ addiert und von $\boldsymbol{\Theta}_a$ subtrahiert, so dass die entstehenden Matrizen positiv definit ist, d.h. wir definieren die Zerlegung

$$\boldsymbol{\Sigma} = \underbrace{(\boldsymbol{\Phi}_a + \boldsymbol{U})}_{=:\boldsymbol{\Phi}_b} + \underbrace{(\boldsymbol{\Theta}_a - \boldsymbol{U})}_{=:\boldsymbol{\Theta}_b} = \boldsymbol{\Phi}_b + \boldsymbol{\Theta}_b \tag{5.16}$$

Die Modelle Ma und Mb besitzen denselben Modellfit, allerdings verschiedene Reliabilitäten ρ_a und ρ_b, denn

$$\rho_a = \frac{\operatorname{tr}(\boldsymbol{J}\boldsymbol{\Phi}_a)}{\operatorname{tr}(\boldsymbol{J}\boldsymbol{\Sigma})} \quad \text{bzw.} \quad \rho_b = \frac{\operatorname{tr}(\boldsymbol{J}\boldsymbol{\Phi}_a) + \operatorname{tr}(\boldsymbol{J}\boldsymbol{U})}{\operatorname{tr}(\boldsymbol{J}\boldsymbol{\Sigma})} \tag{5.17}$$

Dabei ist $\boldsymbol{J} = \mathbf{1}\mathbf{1}'$ die Matrix, die aus lauter Einsen besteht. Die Reliabilität ρ_a des Modells Ma ist also durch die Summe $\operatorname{tr}(\boldsymbol{J}\boldsymbol{U})$ der Einträge in der Matrix $\boldsymbol{U}$ veränderbar. Fällt diese Summe positiv aus, so hat das Modell Mb eine höhere Reliabilität als Modell Mb. Fällt diese Summe negativ aus, so sinkt die Reliabilität. Dies ist aber dazu äquivalent, dass die mittlere Residualkovarianz in Modell Mb gegenüber Modell Ma erhöht wird.

Auf das illustrative Datenbeispiel übertragen heißt dies, dass das Testlet-Modell M2 die Darstellung $\boldsymbol{\Sigma} = \boldsymbol{\Phi}_2 + \boldsymbol{\Theta}_2$ besitzt. Dabei ist $\boldsymbol{\Theta}_2$ eine Diagonalmatrix und die Matrix $\boldsymbol{\Phi}_2$ lässt sich schreiben $\boldsymbol{\Phi}_2 = \tau_\theta \mathbf{1}\mathbf{1}' + \boldsymbol{\Phi}_u$ mit der Matrix $\tau_\theta \mathbf{1}\mathbf{1}'$ für den generellen Faktor θ und einer blockdiagonalen Matrix $\boldsymbol{\Phi}_u$ für die Testlet-Faktoren. In Modell M3 wird die Matrix $\boldsymbol{\Phi}_u$ als Bestandteil der Residualkovarianz angesehen Demzufolge ergeben sich für M2 und M3 die Matrixzerlegungen

$$\text{M2} : \boldsymbol{\Sigma} = \boldsymbol{\Phi}_2 + \boldsymbol{\Theta}_2 = \underbrace{\tau_\theta \mathbf{1}\mathbf{1}' + \boldsymbol{\Phi}_u}_{=\boldsymbol{\Phi}_2} + \qquad \boldsymbol{\Theta}_2 \tag{5.18}$$

$$\text{M3} : \boldsymbol{\Sigma} = \boldsymbol{\Phi}_3 + \boldsymbol{\Theta}_3 = \underbrace{\tau_\theta \mathbf{1}\mathbf{1}'}_{=\boldsymbol{\Phi}_3} + \underbrace{\boldsymbol{\Phi}_u + \boldsymbol{\Theta}_2}_{=\boldsymbol{\Theta}_3} \tag{5.19}$$

Unterschiede in der Reliabilität zwischen Modell M2 und M3 sind demzufolge auf Unterschiede in $\mathbf{1}'\boldsymbol{\Phi}_u\mathbf{1}$ zurückzuführen.

Aufgrund dieser Überlegungen definiert für Westfall und Kollegen (2012) nicht ein Faktormodell (und damit der Modellfit) die Reliabilität, sondern ist umgekehrt ein Faktormodell erst nach Annahme einer Reliabilität definiert. Mit der mittleren Kovarianz $\phi = \operatorname{tr}(\boldsymbol{H}\boldsymbol{\Sigma})/[I(I-1)]$ lässt sich schreiben

$$\boldsymbol{\Sigma} = \phi \mathbf{1}\mathbf{1}' + \boldsymbol{\Theta}^* \quad \text{mit} \quad \boldsymbol{\Theta}^* = \boldsymbol{\Sigma} - \phi \mathbf{1}\mathbf{1}' \tag{5.20}$$

Für (5.20) ergibt sich als Reliabilität Cronbachs Alpha

$$\rho = \alpha = \frac{\phi I^2}{\operatorname{tr}(\boldsymbol{J}\boldsymbol{\Sigma})} \tag{5.21}$$

Wir bemerken, dass das Modell (5.20) äquivalent zur Anpassung eines Modells $\boldsymbol{\Sigma} = \phi\mathbf{1}\mathbf{1}' + \boldsymbol{\Theta} + \boldsymbol{\Delta}$ mit einer diagonalen Residualkovarianzmatrix $\boldsymbol{\Theta}$ und einem mittleren Modellfehler von Null ist (d.h. $\mathbf{1}'\boldsymbol{\Delta}\mathbf{1} = 0$).

Passen wir nun allgemeiner auch eine mittlere residualen Kovarianz σ_ϵ^2 an, so betrachten wir eine Zerlegung

$$\boldsymbol{\Sigma} = \phi\mathbf{1}\mathbf{1}' + \sigma_\epsilon \boldsymbol{H} + \boldsymbol{\Theta} + \boldsymbol{\Delta} \tag{5.22}$$

mit einer Diagonalmatrix $\boldsymbol{\Theta}$. Dieses Modell ist nicht identifiziert, denn man kann nicht zugleich den Anteil der wahren Varianz ϕ und der mittleren Residualkovarianz σ_ϵ schätzen. In Anwendungen gibt man sich daher meistens die lokale stochastische Unabhängigkeit vor, d.h. $\sigma_\epsilon = 0$. Würde man sich jedoch eine Reliabilität ρ_0 vorgeben, dann kann man a priori σ_ϵ so festlegen, dass das Faktormodell (5.22) die Reliabilität ρ_0 besitzt. Die ULS-Anpassung liefert bei einer Fixierung von σ_ϵ

$$\phi_0 = \frac{\mathrm{tr}(\boldsymbol{H}\boldsymbol{\Sigma})}{I(I-1)} - \sigma_\epsilon \tag{5.23}$$

Wählt man

$$\sigma_\epsilon = \frac{\mathrm{tr}(\boldsymbol{H}\boldsymbol{\Sigma})}{I(I-1)} - \frac{\rho_0 \cdot \mathrm{tr}(\boldsymbol{J}\boldsymbol{\Sigma})}{I^2} \tag{5.24}$$

so erhält man als Reliabilität ρ_0. Die Vorgabe einer Reliabilität korrespondiert demzufolge mit der Annahme einer mittleren Residualkovarianz σ_ϵ.

Zuletzt sei noch erwähnt, dass die in einem mehrdimensionalen Modell bestimmte Reliabilität auch alternativ durch ein eindimensionales Modell mit negativ korrelierten Residuen gewonnen werden kann[6]. Danach wäre es durchaus plausibel, die Annahme negativer lokaler stochastischer Abhängigkeiten zu treffen. Während positive lokale Abhängigkeiten einem „Informationsverlust“ gegenüber der Annahme lokaler Unabhängigkeit entsprechen, erhält man einen „Informationsgewinn“ bei negativen lokalen Abhängigkeiten (zu negativen lokalen Abhängigkeiten in IRT-Modellen siehe Habing & Roussos, 2003). Aufgrund äquivalenter statistischer Faktormodelle muss immer eine Identifikationsannahme an die mittlere Residualkovarianz getroffen werden. Durch diese (im konkreten Anwendungsfall auch möglicherweise von einem Mittelwert von Null abweichende Annahme) ist die Reliabilität erst durch diese Annahme eindeutig definiert.

[6] Wir gehen dazu von einem mehrdimensionalen Modell $\boldsymbol{\Sigma} = \boldsymbol{\Phi} + \boldsymbol{\Theta}$ aus. Dann beträgt die Reliabilität dieses Modells $\rho_0 = \mathrm{tr}(\boldsymbol{J}\boldsymbol{\Phi})/\mathrm{tr}(\boldsymbol{J}\boldsymbol{\Sigma})$. Wir untersuchen, welche mittlere Residualkovarianz σ_ϵ aus (5.24) bei Anpassung eines approximativen Einfaktormodells (5.22) mit vorgegebener Korrelation ρ_0 entsteht. Setzt man ρ_0 in (5.24) ein, so folgt

$$\sigma_\epsilon = \frac{\mathrm{tr}(\boldsymbol{H}\boldsymbol{\Phi})}{I(I-1)} - \frac{\mathrm{tr}(\boldsymbol{J}\boldsymbol{\Phi})}{I^2} = \frac{\mathrm{tr}(\boldsymbol{H}\boldsymbol{\Phi}) - (I-1)\mathrm{tr}(\boldsymbol{\Phi})}{I(I-1)} \tag{5.25}$$

Dies lässt sich weiter vereinfachen gemäß

$$\sigma_\epsilon = \frac{\mathrm{tr}(\boldsymbol{J}\boldsymbol{\Phi}) - I \cdot \mathrm{tr}(\boldsymbol{\Phi})}{I(I-1)} = -\frac{\delta}{I-1} \tag{5.26}$$

Dabei ist $\delta = \mathrm{tr}(\boldsymbol{\Phi}) - 1/I \cdot \mathrm{tr}(\boldsymbol{J}\boldsymbol{\Phi})$ die so genannte *deviance from true-score equivalence* (Lucke, 2005). Typischerweise ist δ in mehrdimensionalen Modellen positiv, so dass σ_ϵ negativ ist. Also kann ein mehrdimensionales äquivalent als ein approximatives Einfaktormodell mit mittleren negativen Residualkovarianzen dargestellt werden.

5.2.5 Zwischenresümee

Zusammenfassend sollten unsere Überlegungen zeigen, dass ein guter Modellfit weder notwendig noch hinreichend für die Bestimmung einer richtigen Reliabilität ist. Die Bestimmung der Reliabilität kann dabei modellbasiert (mit einem Faktormodell oder einem IRT-Modell) oder designbasiert (mit einer geeigneten Sampling-Annahme im Domain Sampling) erfolgen. Anhand unseres illustrativen Datenbeispiels haben wir gezeigt, dass für zwei Modelle mit gleichem Modellfit (M2 und M3) verschiedene Reliabilitäten resultieren und daher nicht empirisch die „richtige Reliabilität“ bestimmbar ist. Im Domain Sampling wird die Reliabilität mit der Annahme eines Item Sampling Designs definiert, weshalb die Untersuchung der Passung einzelner Items in einem Faktormodell hier keine Rolle spielt. Allerdings wurde deutlich, dass sich die Annahmen des Domain Samplings auch als ein eindimensionales Faktormodell mit korrelierten Fehlern umschreiben lassen. Typischerweise wird man jedoch in Anwendungen nicht die komplette komplexe Fehlerstruktur in Residuen modellieren und daher Modelle nur approximativ anpassen, so dass ein Modellfehler entsteht. Es wurde gezeigt, dass für eine unverzerrte Reliabilitätsschätzung nur Annahmen an den mittleren Modellfehler getroffen werden müssen und keine perfekte Modellpassung vorliegen muss. Mit dieser Argumentation kann der Einsatz fehlspezifizierter eindimensionaler Modelle begründet werden, wenn diese die „primäre Dimension richtig“ identifizieren.

5.3 HAMLET-Test

Im Folgenden soll anhand eines empirischen Datenbeispiels die Anwendung verschiedener Skalierungsmodelle illustriert werden. Es werden für einen Test zur Erfassung der Lesekompetenz die unterschiedlichen Interpretationen der Reliabilität sowie der Item- und Personenparameter diskutiert. Dabei soll insbesondere auf die Bedeutung der Parameter für individualdiagnostische Fragen eingegangen werden.

5.3.1 Material

Es wurde zur Erfassung der Lesekompetenz bei Grundschülern der HAMLET-Test eingesetzt, der drei Texte mit jeweils vier Items umfasst (Lehmann, Peek & Poerschke, 2006). Die drei Lesetexte können den Textsorten literarischer Text, Sachtext und diskontinuierlicher Text zugeordnet werden.

Ein Testheft mit den drei ausgewählten Lesetexten wurde in einer Teilstichprobe von $N = 328$ österreichischen Viertklässlern im Rahmen einer Bildungsstandardserhebung im Mai 2010 mit einer Testzeit von 20 Minuten vorgelegt. Die Lösungshäufigkeiten der Items variierten zwischen $p = .46$ und $p = .93$ ($M = .72$). Die interne Konsistenz betrug $\alpha = .68$. Das auf Basis der Testlets definierte stratifizierte Cronbachs Alpha fiel mit .70 etwas höher aus. Für die Summenscores der drei Testlets wurde ein Cronbachs Alpha von .56 ermittelt.

5.3.2 Statistische Analysen

Die Datenaufbereitung und alle statistischen Analysen wurden in der Software R (R Core Team, 2014) durchgeführt. Das Rasch-Modell (Modell M1) und das Rasch-Testlet-Modell (Modell M2) wurden mit der Methode Marginal Maximum Likelihood (Adams & Wu, 2007) im R-Paket TAM (Kiefer, Robitzsch & Wu, 2015) geschätzt. Für die Schätzung eines Item-Response-Modells mit korrelierten Residuen (sog. *marginale Modelle*) wurde das R-Paket sirt (Robitzsch, 2015) eingesetzt. Zur Spezifikation der korrelierten Residuen wurde ein Rasch-Copula-Modell mit der boundary mixture Copula verwendet (Braeken, 2011; siehe auch Braeken et al., 2007; Braeken, Kuppens, De Boeck & Tuerlinckx, 2013; Schroeders, Robitzsch & Schipolowski, 2014). In diesem Modell werden die korrelierten Residuen durch die Annahme einer Mischverteilung aus lokal stochastisch unabhängigen und maximal abhängigen Items in einem Testlet modelliert.

5.3.3 Ergebnisse

Modellvergleich

Die Deviance und Informationskriterien der drei Modelle M1, M2 und M3 sind in Tabelle 5.1 dargestellt. Ein Vergleich der Modelle anhand des Informationskriteriums AIC ergab, dass das Testlet-Modell M2 (AIC=3907) und das Copula-Modell M3 (AIC=3894) besser als das Rasch-Modell M1 (AIC=3965) passt. Für die beiden anderen Informationskriterien BIC und CAIC zeigte sich dasselbe Bild. Ein entsprechender Likelihood-Quotienten-Test wurde statistisch signifikant und bestätigte das Befundmuster. Das auf dem AIC basierende Modellgewicht nach Akaike (vgl. Burnham & Anderson, 2004) ermittelt für das Copula-Modell eine klare Modellpräferenz mit dem Gewicht w(M3)=.998.

Tabelle 5.1: *Deviance und Informationskriterien für die Modelle M1, M2 und M3*

	M1	M2	M3
Deviance	3938.71	3874.53	3861.90
#Npar	13	16	16
AIC	3965	3907	3894
BIC	4014	3967	3956
CAIC	4027	3983	3971
w_m	.000	.002	.998

Anmerkungen: M1: Rasch-Modell; M2: Testlet-Modell; M3: Copula-Modell; #Npar: Anzahl geschätzter Parameter; w_m: Akaike-Gewicht für Modell m

Standardabweichung der Faktoren und Reliabilität

In den drei Modellen ergaben sich die folgenden Standardabweichungen für die Traits $SD(\theta)$: 1.09 (M1: Rasch-Modell), 1.09 (M2: Testlet-Modell) und 1.00 (M3: Copula-Modell). Aufgrund der Modellierung der lokalen Abhängigkeiten sank die Standardabweichung im Copula-Modell gegenüber dem Rasch-Modell. Im Testlet-Modell M2 wurden folgende Standardabweichungen für die drei Testlet-Faktoren geschätzt: 0.87 (Testlet A), 0.38 (Testlet B) und 1.60 (Testlet C). Das Testlet C wies demzufolge die höchste, Testlet B

die geringste lokale Abhängigkeit zwischen den Items auf. Dieser Befund steht im Einklang mit dem Copula-Modell M3, in dem die Abhängigkeiten durch einen Parameter δ modelliert werden. Dabei entspricht ein δ-Wert von 0 lokaler Unabhängigkeit der Items innerhalb eines Testlets, ein δ-Wert von 1 maximaler lokaler Abhängigkeit (Braeken, 2011). Auch hier ergab sich für Testlet C die größte Abhängigkeit ($\delta = .30$), es folgten Testlet A ($\delta = .17$) und Testlet B ($\delta = .00$).

Die Reliabilität wurde für die Modelle M1 und M3 über die EAP-Personenschätzer berechnet (Adams, 2005)[7]. Dabei ergab sich für das Rasch-Modell M1 eine Reliabilität von .66, für das Copula-Modell eine niedrigere Reliabilität von .59. Für das Testlet-Modell wurden die Maße ω_h und ω_t auf Basis der geschätzten Parameter im Faktormodell für dichotome Daten nach Green und Yang (2009) ermittelt. Wenn die Testletvarianz als Fehlervarianz interpretiert wird, war die Reliabilität geringer ausgeprägt ($\omega_h = .57$) als wenn sie der Traitvarianz zugeschrieben wird ($\omega_t = .73$). Zur Illustration der praktischen Relevanz dieses Unterschiedes sei angenommen, dass der Summenscore des Lesetests mit einem (perfekt reliablen) Intelligenztest zu $r_{XZ} = .60$ korreliere. Auf Basis der aus den IRT-Modellen gewonnenen Reliabilitätsschätzungen kann eine minderungskorrigierte Korrelation mittels $r_{lat,XZ} = r_{XZ}/\sqrt{Rel(X)}$ berechnet werden. Bei der Verwendung von ω_t ergibt sich eine latente Korrelation von .70. Für ω_h beträgt die latente Korrelation dagegen .79. Die Verwendung der niedrigeren Reliabilitätsschätzung (ω_h) führt somit zu einer höheren latenten Korrelation.

Itemparameterschätzungen

In Tabelle 5.2 sind aus den Testlets A, B und C jeweils zwei Items mit ähnlichen Lösungshäufigkeiten (p-Werten) angegeben. Zusätzlich werden die entsprechenden Schwierigkeitsschätzungen aus den IRT-Modellen M1, M2 und M3 sowie Standardabweichungen des zugehörigen Testlets berichtet.

Tabelle 5.2: *Itemschwierigkeiten und Standardabweichungen der Testleteffekte*

Item	Testlet	p	SD(Testlet)	Rasch (M1)	Testlet (M2)	Copula (M3)
A2	A	.74	0.87	−1.27	−1.41	−1.24
A1		.85		−2.10	−2.31	−2.01
B1	B	.71	0.38	−1.13	−1.15	−1.09
B3		.91		−2.73	−2.78	−2.66
C2	C	.71	1.60	−1.13	−1.42	−1.07
C3		.87		−2.31	−2.95	−2.43

Anmerkungen: p: Lösungshäufigkeit; SD(Testlet): Standardabweichung der Testleteffekte; Rasch, Testlet bzw. Copula bezeichnen die Itemschwierigkeiten in den entsprechend angepassten Modellen

Die Items B1 und C2 besitzen denselben p-Wert von .71 und demzufolge dieselbe Itemschwierigkeit $b_i = -1.13$ im Rasch-Modell (M1). Im Testlet-Modell M2 fällt allerdings auf, dass die Schwierigkeit von Item C2 im Testlet C mit -1.42 deutlich geringer ist

[7] Ist $Var(\theta)$ eine geschätzte Traitvarianz und $Var(\hat{\theta}_{EAP})$ die empirische Varianz der EAP-Personenschätzer, dann ist die EAP-Reliabilität definiert durch $Var(\hat{\theta}_{EAP})/Var(\theta)$.

als für Item B1 mit -1.15. Dies ist darauf zurückzuführen, dass Itemparameter im Testlet-Modell nicht mehr wie im Rasch-Modell *marginal* interpretiert werden dürfen. D.h. die Schwierigkeit wird im Testlet-Modell nicht bezüglich der Definition $logit\, P(X_i = 1|\theta)$, sondern bedingt auf den Testlet-Faktor u_t bezüglich $logit\, P(X_i = 1|\theta, u_t)$ interpretiert. Daraus ergibt sich, dass die im Testlet-Modell gewonnenen Itemschwierigkeiten eine testletspezifische Metrik von $\sqrt{Var(\theta) + Var(u_t)}$ besitzen. D.h. mit größerer Testletvarianz $Var(u_t)$ fallen Itemschwierigkeiten auch tendenziell vom Betrag her höher aus. In einem Testlet-Modell bestimmte Itemparameter können demnach nur lokal innerhalb eines Testlets interpretiert werden und nicht global für den gesamten Test. Aus unserer Sicht ist die lokale Interpretation aber für die diagnostische Praxis nur wenig nützlich, da der Anwender meistens an einer Einschätzung der Schwierigkeit eines Items bezüglich des gesamten Tests interessiert sein dürfte.

Itemschwierigkeiten aus marginalen IRT-Modellen wie dem hier verwendeten Copula-Modell (Braeken, 2011) oder IRT-Modellen mit korrelierten Residuen (Renard et al., 2004) besitzen diese unerwünschte Eigenschaft nicht, sondern bleiben marginal interpretierbar, da die lokalen Abhängigkeiten zwischen den Items mittels korrelierter Residuen modelliert werden. Wir präferieren (wenn IRT-Modelle zur Modellierung lokaler Abhängigkeiten eingesetzt werden) daher den Einsatz marginaler Modelle wie Modell M3.

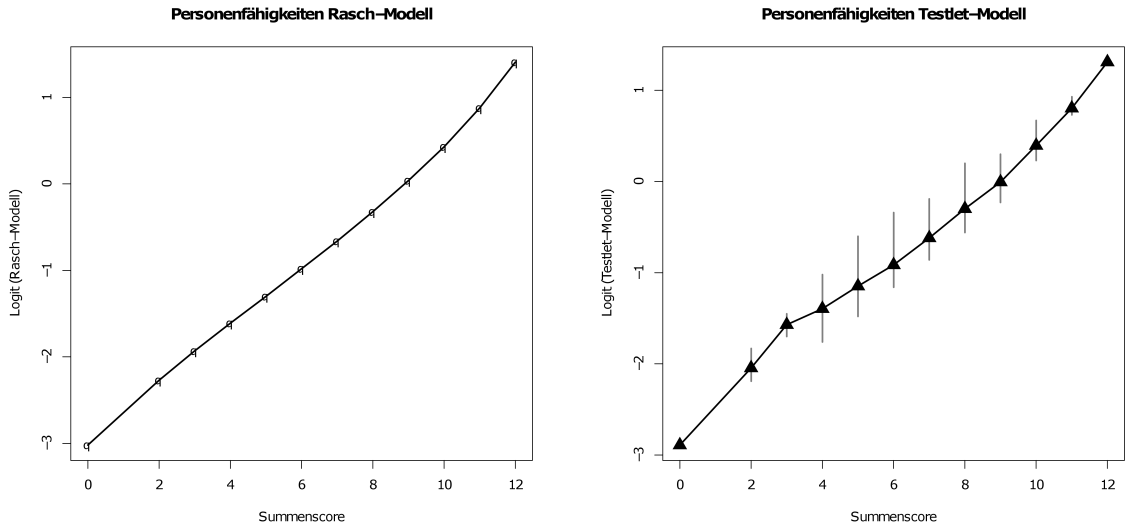

Abbildung 5.3: *Links: Scoring im Rasch-Modell (M1), Rechts: Scoring im Testlet-Modell (M2)*

Personenparameterschätzungen

Im nächsten Schritt werden die aus den IRT-Modellen M1, M2 und M3 gewonnenen Personenparameterschätzungen (EAP-Schätzer) betrachtet. Da wir besonders an den testdiagnostischen Implikationen der verschiedenen Modelle interessiert sind, erscheint es uns angemessen, lediglich 1-PL Modelle zu betrachten, in denen Items dieselbe Diskrimination aufweisen. In der Anwendung würde dies einer Gleichgewichtung der Items entsprechen, die vor allem dann vorgenommen werden sollte, wenn das Testergebnis praktische Konse-

quenzen für die getesteten Personen besitzt (High-stakes Kontext; siehe auch Kolen, 2006, S. 183).

Abbildung 5.3 stellt den Zusammenhang von Summenscore und der Personenparameterschätzung im Rasch-Modell (linke Abbildung) bzw. im Testlet-Modell (rechte Abbildung) dar. Der Summenscore stellt im Rasch-Modelleine suffiziente Statistik für den Personenparameter dar. Die beiden Personenschätzer (Summenscore und Personenparameter aus M1) lassen sich deshalb monoton ineinander transformieren. Für das Testlet-Modell M2 gilt diese Eigenschaft nicht mehr. Die rechte Grafik in Abbildung 5.3 zeigt, dass Personen mit derselben Ausprägung im Summenscore (z.B. Score von 5) verschiedene Personenfähigkeitsschätzungen erhalten (Streuung ist durch vertikale Linie in der Grafik dargestellt). Es wird deutlich, dass im Testlet-Modell M2 sogar die Monotonie-Eigenschaft für die Fähigkeitsschätzer verletzt ist. So weisen z.B. einige Personen mit einem Summenscore von 5 eine höhere Fähigkeitsschätzung im Testlet-Modell auf als Personen mit einem Summenscore von 6 (siehe z.B. Person 21 vs. Person 5). In diesem Sinne verliert man also mit dem Rasch-Testlet-Modell die Eigenschaft der Suffizienz des Summenscores. Bei diagnostischen Entscheidungen (insbesondere im Fall von High-stakes Tests) kann eine solche Scoring-Eigenschaft fragwürdig sein.

Tabelle 5.3: *Korrelationen (R_s) und nicht aufgeklärte Varianz ($\sqrt{1-R_s^2}$) der verschiedenen Schätzer der Personenfähigkeiten*

R_s	Roh	Rasch	Testlet
Rasch (M1)	1.000		
Testlet (M2)	.985	.985	
Copula (M3)	.991	.991	.994

$\sqrt{1-R_s^2}$	Roh	Rasch	Testlet
Rasch (M1)	.000		
Testlet (M2)	.172	.172	
Copula (M3)	.134	.134	.109

Anmerkungen: Roh: Rohwert (Summenscore); Rasch: EAP-Personenschätzer in Rasch-Modell (M1); Testlet: EAP-Personenschätzer im Testlet-Modell (M2); Copula: EAP-Personenschätzer in Copulamodell (M3)

Abschließend soll untersucht werden, wie stark sich die aus den drei Modellen resultierenden Personenfähigkeitsschätzer unterscheiden. Zur Bestimmung der Ähnlichkeit wird die Spearman-Korrelation R_s verwendet. Die Spearman-Korrelation besitzt gegenüber der Pearson-Korrelation den Vorteil, dass sie lediglich auf die Rangreihe zurückgreift. Da viele Autoren (siehe Lord, 1980; Ramsay, 1996) den aus IRT-Modellen resultierenden Personenparameterschätzern nur Bedeutung im Hinblick auf die Invarianz von Rangreihenfolgen zuschreiben (also Ordinalskalenniveau), bevorzugen wir die Spearman-Korrelation gegenüber der Pearson-Korrelation. Des Weiteren erscheint es uns für die Interpretation des Ausmaßes der Nichtübereinstimmung hilfreich die „Genauigkeit“ von Personenparametern in der Metrik individueller Standardmessfehler zu berichten. Demzufolge verwenden wir zusätzlich als Maß für die Nichtübereinstimmung die nicht aufgeklärte Varianz bzw. Streuung $\sqrt{1-R_s^2}$. In Tabelle 5.3 sind die Korrelationen und nicht aufgeklärten Varianzen abgetragen. Es wird deutlich, dass der Schätzer aus dem Rasch-Modell etwas niedriger mit dem Testlet-Modell ($R_s = .985$) als mit dem Copula-Modell ($R_s = .991$) korreliert. Personenschätzungen aus dem Copula-Modell weisen einen hohen Zusammenhang mit denen des Testlet-Modells auf ($R_s = .994$). Allerdings bleibt zwischen beiden Schätzern noch ein Unterschied von $\sqrt{1-R_s^2} = .109$ bestehen, was etwa einer Zehntel Standardabweichung

in der Metrik des individuellen Standardmessfehlers entspricht.

Tabelle 5.4: *Vergleich der Personenparameterschätzungen (EAP) aus verschiedenen Modellen*

Person	Summe	PABC	Rasch	Copula	Testlet	Testlet A	Testlet B	Testlet C
1	5	$P014$	−1.30	−1.28	−1.48	−0.82	−0.12	1.51
2	5	$P113$	−1.30	−1.20	−1.36	−0.34	−0.11	0.27
3	5	$P023$	−1.30	−1.17	−1.24	−0.88	0.00	0.21
4	5	$P212$	−1.30	−1.13	−1.24	0.10	−0.13	−0.78
5	6	$P114$	−0.98	−0.96	−1.16	−0.42	−0.15	1.34
6	5	$P122$	−1.30	−1.05	−1.13	−0.42	−0.01	−0.84
7	6	$P213$	−0.98	−0.90	−1.08	0.04	−0.14	0.09
8	6	$P024$	−0.98	−0.94	−1.05	−0.95	−0.02	1.24
9	6	$P123$	−0.98	−0.85	−0.96	−0.49	−0.02	0.02
10	5	$P320$	−1.30	−1.10	−0.89	0.47	−0.02	−2.89
11	7	$P214$	−0.66	−0.69	−0.86	−0.04	−0.17	1.20
12	6	$P033$	−0.98	−0.84	−0.85	−1.02	0.09	−0.03
13	6	$P222$	−0.98	−0.79	−0.84	−0.04	−0.04	−1.03
14	7	$P313$	−0.66	−0.71	−0.79	0.42	−0.17	−0.08
15	5	$P230$	−1.30	−1.00	−0.76	−0.10	0.08	−3.03
16	7	$P124$	−0.66	−0.61	−0.74	−0.56	−0.05	1.13
17	6	$P132$	−0.98	−0.73	−0.74	−0.56	0.09	−1.08
18	7	$P223$	−0.66	−0.59	−0.67	−0.11	−0.05	−0.15
19	7	$P412$	−0.66	−0.66	−0.65	0.88	−0.18	−1.16
20	7	$P034$	−0.66	−0.60	−0.62	−1.10	0.08	1.05
21	5	$P140$	−1.30	−0.93	−0.60	−0.70	0.19	−3.39
22	6	$P231$	−0.98	−0.72	−0.60	−0.11	0.07	−2.10
23	7	$P322$	−0.66	−0.52	−0.56	0.34	−0.06	−1.22
24	7	$P133$	−0.66	−0.52	−0.55	−0.63	0.06	−0.21
25	7	$P232$	−0.66	−0.47	−0.44	−0.18	0.06	−1.28
26	7	$P043$	−0.66	−0.50	−0.44	−1.14	0.17	−0.26
27	6	$P240$	−0.98	−0.71	−0.34	−0.26	0.16	−3.32
28	7	$P331$	−0.66	−0.50	−0.30	0.25	0.05	−2.31
29	7	P241	−0.66	−0.41	−0.19	−0.24	0.17	−2.37

Anmerkungen: Summe: Summenscore; PABC: Antwortmuster der Summenscores der Testlets A, B und C; Rasch: EAP Rasch-Modell; Copula: EAP Copula-Modell; Testlet: EAP Trait Testlet-Modell; Testlet A, B, C: EAP für Testleteffekte aus Testlet-Modell

Anhand eines Ausschnitts aus dem Datensatz sollen die Unterschiede der Personenparameterschätzer zwischen den drei Modellen illustriert werden (siehe Tabelle 5.4 auf S. 109). Es sind für 29 Personen mit jeweils einem Summenscore von 5, 6 oder 7 Punkten die verschiedenen Schätzer angegeben. Dabei fällt beispielsweise auf, dass unter Personen mit einem Summenscore von 5 diejenigen mit hohen Scores auf dem Testlet C die niedrigsten Scores im Testlet-Modell erhalten (z.B. Person 1 mit Muster $P014$ erhält -1.48, Person 2 mit $P114$ erhält -1.36). Personen mit hohen Scores auf Testlet B (Person 15, Muster $P230$ erhält -0.76) erhalten relativ hohe Fähigkeitsschätzungen. Dieser Befund kann darauf zurückgeführt werden, dass bei „Marginalisierung“ (Ausintegration) der Testletfaktoren Items mit hohen Testletvarianzen niedriger gewichtet werden (siehe Ip,

2010). Da die Testletvarianz für Testlet C am höchsten, für Testlet B am geringsten ausfiel, besitzt ein richtig gelöstes Item im Testlet B demzufolge eine größere Bedeutung für den im Testlet-Modell bestimmten Fähigkeitswert als ein richtiges Item im Testlet C. Es kann kritisch hinterfragt werden, ob dieses empirisch aufgrund der Höhe der Abhängigkeiten bestimmte Scoring valider ist als ein alternativ definiertes Scoring, das z.B. auf der Gleichgewichtung der Items beruht (siehe Brennan, 2001b). Aus unserer Sicht ist es nicht ausreichend, die Entscheidung für ein bestimmtes Scoring durch eine bessere Passung des faktorenanalytischen Modells (z.B. des Testlet-Modells) zu begründen.

5.4 Bedeutung der lokalen Abhängigkeit im Kontext des Reliabilität-Validitäts-Dilemmas

5.4.1 Ein gemeinsames Modell für Reliabilität und Validität

Die bisherigen Überlegungen konzentrierten sich auf verschiedene Ansätze zur Bestimmung und Interpretation der Reliabilität eines Tests mit Testletstruktur. Im Folgenden soll die Perspektive erweitert werden und neben der Reliabilität auch die Validität in die Betrachtungen miteinbezogen werden. Während die Reliabilität die Messgenauigkeit eines Tests angibt, bezieht sich die Validität auf das zu erfassende Konstrukt. Danach misst ein valider Test das, was er zu messen vorgibt. Das Zusammenspiel der Reliabilität und Validität soll anhand eines Lesetests zur Erfassung der Lesekompetenz diskutiert werden.

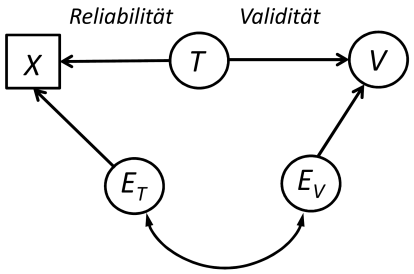

Abbildung 5.4: *Messmodell für Reliabilität und Validität*

Zunächst soll kurz das Grundproblem skizziert werden. Angenommen es soll durch einen Lesetest X die Lesekompetenz V erfasst werden, wobei mit T die wahren Werte im Lesetest bezeichnet seien. Diese Konstellation ist in Abbildung 5.4 als Pfaddiagramm dargestellt. Die Messfehler E_T geben den Anteil in den beobachteten Werten X an, die unabhängig von den wahren Werten T sind. Des Weiteren wird durch den Fehler E_V zum Ausdruck gebracht, dass auch die Lesekompetenz V nicht perfekt durch die Lesefähigkeit T vorhergesagt wird. Wir nehmen an, dass für die Definition von X der *klassische Messfehler* (Carroll, Ruppert, Stefanski & Crainiceanu, 2012) gilt

$$X = T + E_T \quad \text{mit} \quad Cov(T, E_T) = 0 \tag{5.27}$$

Für die mit der Validität assoziierte Variable V gilt die Definitionsgleichung

$$V = T + E_V \quad \text{mit} \quad Cov(T, E_V) = 0 \tag{5.28}$$

Die Variable T ist demzufolge im Hinblick auf V mit einem sog. *Berkson-Fehler* (Carroll et al., 2012) behaftet[8]. Würde X eine perfekt reliable Messung von T darstellen, so gilt $V = X + E_V$. Daraus wird ersichtlich, dass selbst bei einem perfekt reliablen Test noch eine Inferenz von X auf die unbekannte Größe V vorgenommen wird. Der Berkson-Fehler - bringt im Gegensatz zum klassischen Fehler - besser zum Ausdruck, dass die Vorhersage der unbekannten Größe V immer mit einem Fehler E_V behaftet sein wird, auch wenn der Test eine perfekte Reliabilität aufweist. Es soll zusätzlich betont werden, dass ausschließlich die Variable X beobachtbar ist. Für Aussagen zu Zusammenhängen von X mit T bzw. V müssen Annahmen getroffen werden.

Es können nun zwei Fälle unterschieden werden. Im ersten Fall wird von der Annahme unkorrelierter Fehler ausgegangen (E_T und E_V sind unkorreliert). So könnte beispielsweise nur ein sehr kurzer Lesetest eingesetzt worden sein, der eine geringe Reliabilität aufweist. Die Korrelation zwischen X und V unterschätzt deshalb die Validität des Tests bzw. die Korrelation zwischen T und V. Mit Hilfe der klassischen Formel zur Minderungskorrektur kann unter der Annahme unkorrelierter Fehler die Korrelation zwischen X und V korrigiert werden, so dass eine unverzerrte Schätzung der Validität möglich ist.

Im zweiten Fall wird nicht mehr davon ausgegangen, dass die Fehler unkorreliert sind. In dieser Konstellation könnte die Korrelation zwischen X und V auch höher ausfallen als der Zusammenhang zwischen T und V. Dieser Fall ist in der Literatur auch als „Reliabilitäts-Validitätsdilemma“ bekannt (Feldt, 1997; Rost, 2004) und bezeichnet das Phänomen, dass eine Erhöhung der Reliabilität eines Tests häufig mit einer Verringerung der Validität verbunden ist. Das Dilemma unterstreicht die Notwendigkeit die Reliabilität eines Tests nicht losgelöst von seiner Validität zu betrachten. So könnte z.B. bei einem Lesetest auf komplexe Textstimuli mit mehreren Aufgaben verzichtet werden, um die Abhängigkeit zwischen den Testaufgaben zu reduzieren (Erhöhung der Reliabilität). Dies würde aber gleichzeitig zur Konsequenz haben, dass der Test aufgrund des stark vereinfachten Aufgabenmaterials weniger geeignet zur Erfassung eines so komplexen Merkmals wie der Lesekompetenz wäre (Verringerung der Validität).

Im Folgenden sollen die Überlegungen etwas stärker formalisiert werden (vgl. Feldt, 1997). Wir nehmen an, dass die Testlet-Faktoren U und der generelle Faktor θ Bestandteil der wahren Varianz seien. Dann können wir für die Variable T des wahren Wertes annehmen

$$X = \theta + U + E_T \quad \text{mit} \quad T = \theta + U \tag{5.29}$$

[8]Während sich für eine messfehlerbehaftete Variable X nach dem klassischen Messfehler die Gleichung $X = T + E$ mit unkorrelierten Variablen T und E ergibt, so gilt nach dem Berkson-Fehler $T = X + E$ mit unkorrelierten Variablen X und E. Der Berkson-Fehler entspricht demzufolge einem formativen Messfehler mit einer Fehlervariablen (Diamantopoulos, 2006). Demzufolge gilt im klassischen Messmodell $Var(X) > Var(T)$ und damit $|Cor(X, Z)| < |Cor(T, Z)|$ für eine Kovariate Z, während im Modell mit Berkson-Fehler $Var(X) < Var(T)$ und $|Cor(X, Z)| > |Cor(T, Z)|$ gilt. Betrachtet man die Regression von X auf Z (`X ~ Z`), so ist der Regressionskoeffizient im klassischen Fehlermodell erwartungstreu, im Modell mit Berkson-Fehler jedoch verzerrt. Vertauscht man die Rollen von X und Z und betrachtet die Regression von Z auf X (`Z ~ X`), so ist der Regressionskoeffizient im klassischen Fehlermodell verzerrt, im Modell mit Berkson-Fehler jedoch erwartungstreu (siehe Carroll et al., 2012). Bei Personenparameterschätzungen in Messmodellen folgt der Maximum-Likelihood-Schätzer (MLE oder WLE) dem klassischen Fehlermodell, der EAP dem Berkson-Fehler.

wobei die Variablen θ, U und E_T unkorreliert sind. Die Validität V ist definiert als

$$V = T + E_V = \theta + U + E_V \tag{5.30}$$

mit unkorrelierten Variablen T und E_V.

Die als Korrelation des Testwertes X mit der unbekannten Lesekompetenz V definierte *Validität* $Val(X)$ kann bestimmt werden durch

$$\begin{aligned} Val(X) \;=\; Cor(X,V) &= \frac{Var(T)}{\sqrt{Var(T)+Var(E_T)}\sqrt{Var(T)+Var(E_V)}} \\ &= \sqrt{\frac{Var(T)}{Var(T)+Var(E_T)}} \cdot \sqrt{\frac{Var(T)}{Var(T)+Var(E_V)}} \\ &= \sqrt{Rel_T(X)}\sqrt{Val_{lat,T}(X)} \end{aligned} \tag{5.31}$$

Mit $Val_{lat,T}(X) = Cor(T,V)$ wird die *latente Validität* (minderungskorrigierte Validität) bezeichnet. Aus dieser Formel wird ersichtlich, dass die Reliabilität $Rel_T(X)$ eine notwendige Voraussetzung für die als Korrelation $Val(X) = Cor(X,V)$ berechnete Validität ist. Wenn die latente Validität $Val_{lat,T}(X)$ festgehalten wird, sinkt die beobachtete Validität mit abnehmender Reliabilität des Tests. Diese Interpretation ist allerdings nur gültig, wenn E_T und E_V unkorreliert sind. Des Weiteren gilt es zu beachten, dass die Größen $Rel_T(X)$ und $Val_{lat,T}(X)$ im Gegensatz zu $Val(X)$ von der Definition des wahren Wertes T im Test abhängen.

Alternativ könnte nun die Varianz der wahren Werte dadurch definiert werden, dass die Testletvarianz dem Fehler zugeschrieben wird. Daraus würden sich folgende Variablen ergeben: $\tilde{T} = \theta$, $\tilde{E}_{\tilde{T}} = E_T + U$, $\tilde{E}_V = E_V + U$.Dann sind die Fehlervariablen $\tilde{E}_{\tilde{T}}$ und $\tilde{E}_V$ positiv korreliert, denn

$$Cov(\tilde{E}_{\tilde{T}}, \tilde{E}_V) = Cov(U + E_T, U + E_V) = Var(U) > 0 \tag{5.32}$$

Auf das Beispiel der Lesekompetenz übertragen hieße dies, dass bei der Zuschreibung der testletspezifischen Varianz zum Fehler dieser Varianzanteil zugleich in der Validität V fehlt und demzufolge die beiden Fehler $\tilde{E}_{\tilde{T}}$ und $\tilde{E}_V$ positiv korreliert sein müssen.

Wie wirkt sich die veränderte Definition des wahren Wertes nun auf das Verhältnis von Reliabilität und Validität aus? Wegen $Var(T) = Var(\theta) + Var(U)$ folgt die Beziehung

$$\begin{aligned} Val(X) &= \frac{Var(\theta)+Var(U)}{\sqrt{Var(\theta)+Var(U)+Var(E_T)} \cdot \sqrt{Var(\theta)+Var(U)+Var(E_V)}} \\ &= \frac{\sqrt{Var(\theta)}}{\sqrt{Var(\theta)+Var(U)+Var(E_T)}} \cdot \frac{\sqrt{Var(\theta)}}{\sqrt{Var(\theta)+Var(U)+Var(E_V)}} + \\ &\qquad \frac{Var(U)}{\sqrt{Var(\theta)+Var(U)+Var(E_T)} \cdot \sqrt{Var(\theta)+Var(U)+Var(E_V)}} \\ &= \sqrt{Rel_{\tilde{T}}(X)} \cdot \sqrt{Val_{lat,\tilde{T}}(X)} + \frac{Cov(\tilde{E}_{\tilde{T}}, \tilde{E}_V)}{SD(X)SD(V)} \end{aligned} \tag{5.33}$$

Für die beobachtete Validität wird folgendes aus dieser Beziehung deutlich. Erstens hängt die beobachtete Validität auch von den korrelierten Fehler ab. Obwohl die Reliabilität $Rel_{\tilde{T}}(X)$ geringer ausfallen wird, da $Var(U)$ nicht mehr der wahren Varianz zugeordnet wird, muss die Validität nicht zwangsläufig sinken, da aufgrund der korrelierten Fehler der rechte Term in Gleichung (5.33) positiv sein wird. Zweitens besitzen die korrelierten Fehler Konsequenzen für die Minderungskorrektur der beobachteten Validität. So könnte eine Adjustierung der Validität anhand der Reliabilität $Rel_{\tilde{T}}(X)$ mit folgender Formel vorgenommen werden:

$$Val_{lat,\tilde{T}}(X) = \frac{[Val(X)]^2}{Rel_{\tilde{T}}(X)} \tag{5.34}$$

Die Korrelation der Fehler $\tilde{E}_{\tilde{T}}$ und $\tilde{E}_V$ wird in dieser Formel allerdings nicht berücksichtigt. Dies wird im Allgemeinen zu einer Überschätzung der Validität durch die Minderungskorrektur führen.

5.4.2 Einschränkung der Validität durch Elimination lokaler Abhängigkeiten

Im Folgenden wird davon ausgegangen, dass die Testletvarianz $Var(U)$ als „Störfaktor“ betrachtet wird. Damit verbunden wird das Ziel verfolgt, die Testletvarianz in den konkreten Messungen zu eliminieren. Wie wirkt sich die Elimination der Testletvarianz auf die Reliabilität und Validität des Tests aus? Dazu soll anstelle der Messung $X = \theta + U + E_T$ mit $E_T = E_\theta + E_U$ (E_θ und E_U seien unkorreliert) die um den Testleteffekt „bereinigte“ Messung $X^* = \theta + E_\theta$ betrachtet werden. Es wird untersucht, wie sich die Validität des neuen Tests $Val(X^*) = Cor(X^*, V)$ von der Validität des ursprünglichen Tests $Val(X) = Cor(X, V)$ unterscheidet. Wir führen dafür die Abkürzungen σ^2_θ usw. für die entsprechenden Varianzen ein. Die Reliabilitäten der Testkomponenten θ und U sind dann definiert als

$$r_\theta = Rel(\theta) = \frac{\sigma^2_\theta}{\sigma^2_\theta + \sigma^2_{E_\theta}} \quad \text{bzw.} \quad r_U = Rel(U) = \frac{\sigma^2_U}{\sigma^2_U + \sigma^2_{E_U}} \tag{5.35}$$

Für den ursprünglichen Test $X = T + E = \theta + U + E_\theta + E_U$ beträgt die Reliabilität

$$Rel_T(X) = \frac{\sigma^2_\theta + \sigma^2_U}{\sigma^2_\theta + \sigma^2_U + \sigma^2_{E_\theta} + \sigma^2_{E_U}} \tag{5.36}$$

Die latente Validität für den ursprünglichen Test ist definiert als

$$Val_{lat,T}(X) = \frac{\sigma^2_\theta + \sigma^2_U}{\sigma^2_\theta + \sigma^2_U + \sigma^2_{E_V}} \tag{5.37}$$

Für den Test X^* ergibt sich mit $T^* = \theta$

$$Rel_{T^*}(X^*) = r_\theta \quad \text{sowie} \quad Val_{lat,T^*}(X^*) = \frac{\sigma^2_\theta}{\sigma^2_\theta + \sigma^2_U + \sigma^2_{E_V}} \tag{5.38}$$

Offensichtlich besitzt X^* eine geringere latente Validität als X, denn es ist $Val_{lat,T^*}(X^*) < Val_{lat,T}(X)$. Je größer also der fehlende Varianzanteil σ_U^2 ausfällt, desto geringer ist die Validität des neuen Tests X^*. In den meisten Anwendungen wird jedoch die Reliabilität von θ höher sein als die von U ($r_\theta > r_U$). In diesem Fall lässt sich zeigen, dass die Reliabilität von X^* höher ist als die von X, denn

$$Rel_T(X) = \frac{\sigma_\theta^2 + \sigma_U^2}{\frac{1}{r_\theta}\sigma_\theta^2 + \frac{1}{r_U}\sigma_U^2} < \frac{\sigma_\theta^2 + \sigma_U^2}{\frac{1}{r_\theta}\sigma_\theta^2 + \frac{1}{r_\theta}\sigma_U^2} = r_\theta = Rel_{T^*}(X^*) \tag{5.39}$$

Für die Validität des neuen Tests $Val(X^*) = Cor(X^*, V)$ ergibt sich

$$Cor^2(X^*, V) = Rel_{X^*}(T^*)Val(T^*) = r_\theta \cdot \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_U^2 + \sigma_{E_V}^2} \tag{5.40}$$

Zur Abschätzung betrachten wir das Verhältnis $f_V = \frac{Cor^2(X,V)}{Cor^2(X^*,V)}$ der quadrierten Validitäten und setzen $v_{\theta U} = \sigma_U^2/\sigma_\theta^2$ als Varianzverhältnis von U und θ sowie das Reliabilitätsverhältnis $r_{\theta U} = r_\theta/r_U$:

$$\begin{aligned} f_V = \left[\frac{Val(X)}{Val(X^*)}\right]^2 = \frac{Cor^2(X,V)}{Cor^2(X^*,V)} &= \frac{\sigma_\theta^2 + \sigma_U^2}{\frac{1}{r_\theta}\sigma_\theta^2 + \frac{1}{r_U}\sigma_U^2} \cdot \frac{1}{r_\theta} \cdot \frac{\sigma_\theta^2 + \sigma_U^2}{\sigma_\theta^2} \\ &= \frac{(1 + v_{U\theta})^2}{1 + r_{\theta U} v_{U\theta}} \end{aligned} \tag{5.41}$$

Fällt der Koeffizient f_V größer als 1 aus, so würde der ursprüngliche Test X eine höhere Validität aufweisen als der Test X^*. Daraus lässt sich ableiten, dass der Test X (mit Testletstruktur) genau dann eine höhere Validität als X^* (unter Elimination lokaler Abhängigkeiten) besitzt, wenn

$$r_{\theta U} < 2 + v_{U\theta} \tag{5.42}$$

Dies tritt dann ein, wenn das Verhältnis der Reliabilitäten $r_{\theta U}$ nicht sehr groß ist. Nehmen wir an, dass $v_{U\theta} = .5$ ist, d.h. der allgemeine Faktor θ ist in der unbekannten Lesekompetenz V doppelt so hoch gewichtet wie testletspezifische Faktoren U. Dann folgt $f_V > 1$, falls $r_{\theta U} < 2.5$. Mit anderen Worten: Die Reliabilität von θ muss mindestens 2.5-mal so hoch wie die Reliabilität von U sein, damit der die lokalen Abhängigkeiten eliminierende Test X^* eine höhere Validität als der ursprüngliche Test X besitzt.

Optimale Gewichtung der Testkomponenten im Hinblick auf Validität

Es wurde gezeigt, wie sich die vollständige Elimination der Testletvarianz U auf die Validität des Tests X^* auswirkt. Abschließend soll untersucht werden, welche Gewichtung von θ und U zu einem Test X mit maximaler Validität führt. Anstatt die Testletvarianz U als „Störfaktor“ zu entfernen, soll die vorliegende Trait- und Testletvarianz so gewichtet werden, dass die Validität des Tests maximiert wird.

Wir betrachten dazu die Testvariable $X_w = w(\theta + E_\theta) + (1-w)(U + E_U)$ für $0 \leq w \leq 1$. Es soll ein Gewichtungsparameter w bestimmt werden, so dass die quadrierte Validität

$Val(X_w)^2 = Cor^2(X_w, V)$ maximiert wird. Die quadrierte Korrelation lässt sich schreiben als

$$\begin{aligned}[Val(X_w)]^2 = Cor^2(X_w, V) &= \frac{[w\sigma_\theta^2 + (1-w)\sigma_U^2]^2}{w^2(\sigma_\theta^2 + \sigma_{E_\theta}^2) + (1-w)^2(\sigma_U^2 + \sigma_{E_U}^2)} \\ &= \sigma_\theta^2 \cdot \frac{[w + (1-w)v_{U\theta}]^2}{w^2\frac{1}{r_\theta} + (1-w)^2\frac{1}{r_U}v_{U\theta}}\end{aligned} \tag{5.43}$$

Dabei ist die Reliabilität gegeben durch

$$Rel_{T_w}(X_w) = \frac{w^2\sigma_\theta^2 + (1-w)^2\sigma_U^2}{w^2(\sigma_\theta^2 + \sigma_{E_\theta}^2) + (1-w)^2(\sigma_U^2 + \sigma_{E_U}^2)} = \frac{w^2 + (1-w)^2 v_{U\theta}}{w^2\frac{1}{r_\theta} + (1-w)^2\frac{1}{r_U}v_{U\theta}} \tag{5.44}$$

Die Maximierung der Validität (5.43) in Abhängigkeit von w ergibt die optimale Gewichtungskonstante w^* [9]

$$w^* = \frac{r_\theta}{r_\theta + r_U} = \frac{r_{\theta U}}{r_{\theta U} + 1} \tag{5.52}$$

Bei perfekt reliablen Messungen der Testkomponenten θ und U ($r_\theta = 1$ und $r_U = 1$) würde man eine Gleichgewichtung $w^* = 0.5$ vornehmen. Eine Maximierung der Reliabilität würde zu einem Gewicht $w^* = 1$ führen. Unter der Annahme dass die Reliabilität von θ doppelt so groß sei wie die von U (also $r_{\theta U} = 2$), würde die Gewichtungskonstante $w^* = 2/3$ betragen. In Abbildung 5.5 findet sich beispielhaft die Darstellung der Reliabilität und der Validität in Abhängigkeit von w. Es fällt auf, dass es Gewichtungen w gibt, für die die Validität größer als die Reliabilität ist.

[9] Wir betrachten die Testvariable $X_w = w(\theta + E_\theta) + (1-w)(U + E_U)$ für $0 \leq w \leq 1$. Es soll ein Gewichtungsparameter w bestimmt werden, so dass die quadrierte Validität $Val(X_w)^2 = Corr^2(X_w, V)$ maximiert wird. Dann können wir schreiben

$$Corr^2(X_w, V) = f(w) = \frac{[g(w)]^2}{h(w)} = \frac{\left[w\sigma_\theta^2 + (1-w)\sigma_U^2\right]^2}{w^2(\sigma_\theta^2 + \sigma_{E_\theta}^2) + (1-w)^2(\sigma_U^2 + \sigma_{E_U}^2)} \tag{5.45}$$

Dann ist $f' = 0$ genau dann, wenn $2gg'h - g^2h' = 0$, also $2g'h - gh' = 0$. Wir ermitteln

$$g'(w) = \sigma_\theta^2 - \sigma_U^2 \tag{5.46}$$

$$h'(w) = 2w(\sigma_\theta^2 + \sigma_{E_\theta}^2) - 2(1-w)(\sigma_U^2 + \sigma_{E_U}^2) \tag{5.47}$$

Weiter ist

$$\begin{aligned}2g'h &= 2w^2(\sigma_\theta^2 + \sigma_{E_\theta}^2)\sigma_\theta^2 - 2(1-w)^2(\sigma_U^2 + \sigma_{E_U}^2)\sigma_\theta^2 \\ &\quad -2w^2(\sigma_\theta^2 + \sigma_{E_\theta}^2)\sigma_U^2 + 2(1-w)^2(\sigma_U^2 + \sigma_{E_U}^2)\sigma_U^2\end{aligned} \tag{5.48}$$

$$\begin{aligned}gh' &= 2w^2(\sigma_\theta^2 + \sigma_{E_\theta}^2)\sigma_\theta^2 - 2w(1-w)(\sigma_U^2 + \sigma_{E_U}^2)\sigma_\theta^2 \\ &\quad -2w(1-w)(\sigma_\theta^2 + \sigma_{E_\theta}^2)\sigma_U^2 + 2(1-w)^2(\sigma_U^2 + \sigma_{E_U}^2)\sigma_U^2\end{aligned} \tag{5.49}$$

Wir leiten weiter ab

$$\begin{aligned}\frac{2g'h - gh'}{1-w} &= [2w - 2(1-w)]\,(\sigma_U^2 + \sigma_{E_U}^2)\sigma_\theta^2 \\ &\quad + [-2w]\,(\sigma_\theta^2 + \sigma_{E_\theta}^2)\sigma_U^2\end{aligned} \tag{5.50}$$

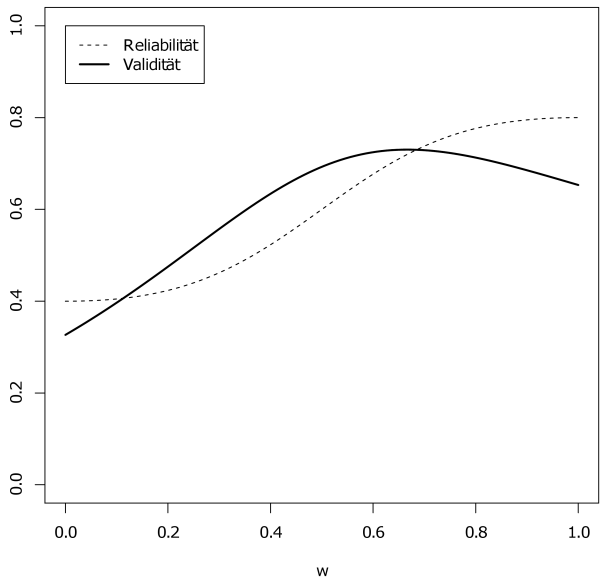

Abbildung 5.5: *Darstellung der Reliabilität und Validität für die Variable X_w mit $v_{U\theta} = .5$, $r_\theta = .8$, $r_U = .4$ und $\sigma^2_\theta = .53$*

5.4.3 Zusammenfassung

In diesem Abschnitt wurde ein Modell zur gemeinsamen Betrachtung der Reliabilität und Validität eines Tests eingeführt. Anhand dieses Modells wurde das klassische Reliabilitäts-Validitätsdilemma diskutiert und gezeigt, welche unterschiedlichen Auswirkungen die Interpretation der Testletvarianz auf die Reliabilität und Validität eines Tests haben kann. Zusätzlich wurde untersucht, welche Konsequenzen mit einer Elimination der Testletvarianz aus dem Test verbunden sind. Während die Reliabilität des Tests durch die Elimination der Testletvarianz ansteigen wird, ist im Allgemeinen davon auszugehen, dass die Validität des so konstruierten Tests geringer sein wird.

Die in diesem Abschnitt untersuchte Größe $Cor(X, V)$ wird in der Generalisierbarkeits-

---

Der optimale Koeffizient w^* ist gegeben durch

$$\begin{aligned} w^* &= \frac{\sigma^2_\theta(\sigma^2_U + \sigma^2_{E_U})}{\sigma^2_\theta(\sigma^2_U + \sigma^2_{E_U}) + \sigma^2_U(\sigma^2_\theta + \sigma^2_{E_\theta})} \\ &= \frac{(\sigma^2_U + \sigma^2_{E_U})}{(\sigma^2_U + \sigma^2_{E_U}) + v_{U\theta}(\sigma^2_\theta + \sigma^2_{E_\theta})} \\ &= \frac{\frac{1}{r_U} v_{U\theta}\sigma^2_\theta}{\frac{1}{r_U} v_{U\theta}\sigma^2_\theta + v_{U\theta}\frac{1}{r_\theta}\sigma^2_\theta} \\ &= \frac{r_\theta}{r_\theta + r_U} \\ &= \frac{r_{\theta U}}{r_{\theta U} + 1} \end{aligned} \tag{5.51}$$

theorie auch als *Generalisierbarkeit* bezeichnet (Kane, 1982; Brennan, 2001a) und umfasst sowohl die Reliabilität $Rel_T(X)$ als auch die latente Validität $Val_{lat,T}(X)$ eines Tests[10]. Aus Sicht der Generalisierbarkeitstheorie sind demzufolge Betrachtungen der Reliabilitätsmaximierung niemals unabhängig von der Validität vorzunehmen. Einige Autoren gehen in ihrer Argumentation sogar so weit, dass ausschließlich die Validität – operationalisiert über ein manifestes Außenkriterium – für die Testkonstruktion herangezogen werden sollte (Ballou, 2009; Cunha, Heckman & Schennach, 2010). Dieses Vorgehen kontrastiert die beiden Herangehensweisen, Itemselektion entweder durch Maximierung interner Konsistenz (Itemtrennschärfen, d.h. der Korrelation von Item X_i und T) oder durch die Maximierung der Itemvalidität (Korrelation von X_i und V; siehe Lord & Novick, 1968; Rost, 2004) vorzunehmen. In der Forschungspraxis und testdiagnostischen Anwendungen scheint es vernünftig beide Aspekte bei der Testkonstruktion zu berücksichtigen. Wie eine mögliche optimale Gewichtung beider Kriterien vorgenommen werden könnte, wurde in Abschnitt 5.4.2 kurz skizziert. In der forschungsmethodischen Literatur zu Tests mit Testletstruktur dominieren bisher vor allem Betrachtungen zur Reliabilität.

5.5 Diskussion

Im vorliegenden Beitrag wird betont, dass wahrer Wert und Fehler nicht ohne Modellannahmen definiert werden können. Für die Schätzung der Reliabilität ist vor allem die

[10] Die Generalisierbarkeit wird im Ansatz von Kane (1982) als ein Sampling-Modell aufgefasst. Im Mittelpunkt steht der Gedanke, dass Items verschiedene Facetten (oder Situationen) abbilden. Eine konkrete Messung entsteht dadurch, dass bestimmte Facetten (z.B. Itemformate oder Textsorten) in einem Test fixiert werden. Ein beobachteter Wert X wird nach dem klassischen Fehlermodell in einen wahren Wert T und einen dazu unkorrelierten Fehler E_T zerlegt, d.h. $X = T + E_T$. Die Reliabilität ist dann definiert als $Rel_T(X) = Var(T)/Var(X)$. Die Variable T beschreibt allerdings nur die wahren Werte im administrierten Test unten den fixierten Facetten. Es wird aber häufig von Interesse sein, von T auf eine Variable V zu generalisieren, die alle möglichen Facetten umfasst und die perfekt valide Messung beschreibt. Kane (1982) nimmt an, dass $T = V + E_V$ mit unkorrelierten Variablen V und E_V gilt. Für die mit der Validität assoziierte Variable V gilt in diesem Ansatz also wiederum ein klassisches Fehlermodell. Konträr dazu haben wir in unserem Messmodell in 5.4.1 ein Modell mit Berkson-Fehler angenommen. Die minderungskorrigierte Validität ist im Sampling-Modell von Kane als $Val_{lat,T}(X) = Var(V)/Var(T)$ definiert. Insgesamt kann man für das Sampling-Modell der Generalisierbarkeit schreiben

$$X = T + E_T = V + E_V + E_T \tag{5.53}$$

Die manifeste Validität bzw. Generalisierbarkeit im Ansatz von Kane (1982) und Brennan (2001) ergibt sich dann als

$$Val(X) = Cor(X, V) = \frac{Var(V)}{Var(V) + Var(E_V) + Var(E_T)} = Val_{lat,T}(X) \cdot Rel_T(X) \tag{5.54}$$

Das Reliabilitäts-Validitäts-Dilemma entsteht in Definition (5.54) dadurch, dass eine Fixierung von Facetten typischerweise dazu führt, dass $Var(E_T)$ kleiner wird und daher die Reliabilität steigt, andererseits dadurch jedoch $Var(E_V)$ größer wird und daher die Validität fällt. Würde man im Extremfall alle Facetten für den Test auswählen, so würden T und V übereinstimmen und Unreliabilität durch E_V und E_T beschrieben. Dieser Test wäre dann perfekt (latent) valide, hätte jedoch eine geringere Reliabilität. Die Generalisierbarkeit bleibt in Modell (5.53) jedoch immer konstant. Siehe Brennan (2001, S. 132ff.) für weitere Details.

theoretische Begründung des Modells und die mit dem Modell verbundenen Annahmen zentral. Die Durchführung von Modelltests besitzt dagegen nur einen begrenzten Nutzen für die Wahl des richtigen Modells zur Schätzung der Reliabilität. Anhand eines Datensatzes wurde gezeigt, dass man mit der Anpassung verschiedener psychometrischer Modelle (M1: eindimensionales Modell unter Ignorierung der Abhängigkeiten, M2: Testlet-Modell und M3: eindimensionales Modell mit positiven lokalen Abhängigkeiten) unterschiedliche Reliabilitäten erhalten kann. Für alle drei Modelle wurden Argumente vorgebracht, weshalb sie zur Bestimmung der Reliabilität gewählt werden könnten. Die Wahl des Modells und die damit verbundene Definition der Varianz der wahren Werte besitzt aber nicht nur Konsequenzen für die Bestimmung der Reliabilität, sondern wirkt sich auch auf die Interpretation der latenten Variablen θ aus. In Modell M3 wird z.B. die Testletvarianz als Fehlervarianz interpretiert und ω_h zur Berechnung der Reliabilität verwendet. Die wahre Varianz ergibt sich somit als geteilte Varianz zwischen Items verschiedener Testlets, da die gemeinsame Varianz innerhalb eines Testlets der Fehlervarianz zugeordnet wird. Demzufolge ist die Fähigkeit θ in M3 als Ausmaß definiert, Items zu verschiedenen Stimuli korrekt zu beantworten. Wenn Items desselben Stimulus gelöst werden, dann kann dies zu einer erhöhten Kovarianz führen, die allerdings als Fehlervarianz interpretiert wird. Dagegen wird bei der Verwendung von Modell M2 und ω_t zur Berechnung der Reliabilität die Testletvarianz der wahren Varianz zugeordnet. Nach dieser Interpretation würde die Fähigkeit θ auch das Ausmaß umfassen, Items desselben Stimulus korrekt zu beantworten. Inhaltlich würde diese Perspektive im Einklang mit einer „Stimulusschwierigkeit“ stehen, die sich daraus ergibt, dass die getesteten Personen zunächst den Stimulus „verstanden“ haben müssen, um dann in einem darauf folgenden Schritt die zugehörigen Items lösen zu können. Genau dieser erste Prozessschritt wird in der Zuschreibung der Testletvarianz zur Fehlervarianz bei ω_h (Modell M3) eliminiert.

Neben dem modellbasierten Ansatz kann jedoch auch ein designbasierter Ansatz zur Begründung von Rebilibilitätsmaßen gewählt werden. Für Cronbachs Alpha wurde unter Verwendung einer Item Sampling (Domain Sampling) Perspektive gezeigt, wie eine designbasierte Definition der Reliabilität vorgenommen werden kann (Cronbach & Shavelson, 2004; siehe auch modellbasierte und designbasierte Ansätze in der Survey-Statistik, Binder & Roberts, 2012)[11]. Es wurde deutlich, dass ein Sampling basierter Ansatz mit der Vorstellung eines Faktormodells vereinbar ist, in dem bestimmte Annahmen über die Struktur der Residualkovarianzmatrix getroffen werden. Diese Annahmen werden typischerweise einfacher sein als die tatsächliche Struktur der Residualkovarianzmatrix (z.B. unkorrelierte Residuen). , Für die psychometrische Praxis spielt deshalb der Gedanke der Approximate Factor Models eine besondere Rolle, da sie Annahmen an nichtmodellierte

[11] Das Reliabilitätsmaß Cronbachs Alpha basiert auf einem sehr einfachen Design, in dem die Items als untereinander beliebig austauschbar angesehen werden. Im Rahmen des Domain Samplings können zur Bestimmung der Reliabilität auch komplexere Designs eingesetzt werden (Cronbach & Shavelson, 2004). Würden in unserem illustrativen Beispiel Items innerhalb der Testlets zufällig aus einer Domäne ausgewählt und die Testlets als fest und somit konstruktinhärent interpretiert werden, so würde das stratifizierte Cronbachs Alpha (Rajaratnam, Cronbach & Gleser, 1965) ein geeignetes Reliabilitätsmaß darstellen, das sein modellbasiertes Analogon in ω_t besitzt. Bei zufällig konstruierten Testlets und einer Interpretation der Testletvarianz als konstruktirrelevant, würde Cronbachs Alpha auf Basis der Scores der Testlets ein geeignetes Reliabilitätsmaß darstellen (siehe Gignac, 2014), das eine ähnliche Bedeutung wie ω_h besitzt.

Residuen und somit auch über „mittlere lokale Abhängigkeit“ umsetzen. Eine unverzerrte Schätzung der Reliabilität ist dabei bereits mit einer Modellanpassung unter der Annahme einer mittleren Residualkorrelation von Null möglich, was eine schwächere Annahme als die der lokalen stochastischen Unabhängigkeit (alle Residualkorrelationen sind gleich Null) darstellt. Gemäß Überlegungen von Westfall et al. (2012) ist jedoch die Schätzung der Reliabilität per Definition immer unbestimmt, so dass die alleinige Ableitung der Reliabilität auf Basis eines Modellfits nicht hinreichend ist.

Des Weiteren wurde in Abschnitt 5.4 argumentiert, dass bei der Diskussion von Tests mit einer Testletstruktur nicht ausschließlich auf die Reliabilität fokussiert werden sollte, sondern es auch die Validität zu berücksichtigen gilt. Gerade bei Tests mit einer Testletstruktur kann es eintreten, dass eine Erhöhung der Reliabilität des Tests (z.B. durch Ausschluss bestimmter Stimuli) mit einer Verringerung seiner Validität einhergeht (Feldt, 1997). Die in psychometrischen Anwendungen zu beobachtende Praxis, die aus Testlet-Modellen gewonnene Reliabilitätsschätzung als Maß für die Güte eines Testwertes zu betrachten, ist somit kritisch zu hinterfragen (siehe z.B. Eckes, 2015a).

Wenn aus substanzieller Sicht beurteilt wird, dass lokale stochastische Abhängigkeit im Hinblick auf das zu operationalisierte Konstrukt bedeutsam sind, so muss das nicht zwangsläufig die Verwendung von Testlet-Modellen nach sich ziehen. Die Konsequenzen ignorierter positiver lokaler stochastischer Abhängigkeiten müssen dabei im Hinblick auf Itemparameter und Traitvarianzen sowie auf Personenparameter unterschieden werden. In der Konzeption der essenziellen Eindimensionalität (Stout, 1990; Zhang & Stout, 1999) werden lokale Abhängigkeiten zwischen Items zugelassen, die jedoch in einem unendlichen langen Test nicht bedeutsam sind. Der wahre Wert θ ist dabei als Grenzwert in einem unendlich langen Test definiert, so dass praktisch ein Item Sampling modelliert wird (siehe auch Ellis & Junker, 1997). Die Existenz von Testleteffekten kann demzufolge zu einem essenziell eindimensionalen Test führen.

Die Definition der Testlänge in einem konkret vorgelegten Test ist im Zusammenspiel mit einer gewählten Schätzmethode zu verstehen. Wenn Items in einem Multi Matrix Sampling Design vorgelegt werden (nicht alle Personen erhalten alle Items), so ist es für eine näherungsweise unverzerrte Schätzung von Itemparametern nicht notwendig, dass jeder Testperson viele Items vorgelegt werden, sondern nur, dass insgesamt viele Items im Matrix Sampling Design existieren. Dabei ist relevant, dass nicht die Maximum-Likelihood-Schätzmethode (Full Information Schätzmethode), sondern Limited Information Schätzmethoden wie Pairwise Conditional Maximum Likelihood (Zwinderman, 1995), Pairwise Marginal Maximum Likelihood (McDonald, 1997; Renard et al., 2004) oder Generalized Estimating Equations (GEE; Spiess & Hamerle, 1996; Spiess, Nagl & Hamerle, 1997) eingesetzt werden. Mit Limited Information Schätzmethoden wird die Information aus paarweisen Kovarianzen zwischen Items verwendet. Lässt man die Anzahl der Items größer werden, dann dominieren in der Kovarianzmatrix die Kovarianzen von Itempaaren in verschiedenen Testlets gegenüber Itempaaren innerhalb eines Testlets. Dies begründet, weshalb sich mit wachsender Itemanzahl die Verzerrung in Itemparametern verringert. Bei der Full Information Maximum Likelihood Schätzung wird dagegen (praktisch) in der Likelihood-Funktion nur die Kovarianzmatrix für einer Person vorgelegten Items verwendet (die typischerweise deutlich weniger Items als die Kovarianzmatrix aller Items beinhaltet). Die dem Trait θ unter der falschen Annahme lokaler stochastischer Unab-

hängigkeit zugeschriebene wahre Varianz fällt demzufolge aufgrund weniger vorgelegter Testlets höher aus als bei der Limited Information Anpassung, da in der kleineren Kovarianzmatrix die Kovarianzen innerhalb von Testlets eine größere Bedeutung besitzen. Dass die Maximum Likelihood Schätzung bei fehlspezifizierten Modellen weniger effizient als Limited Information Schätzmethoden (wie zum Beispiel die in Abschnitt 5.2.1 verwendete ULS-Schätzmethode) ist, wurde für Faktorenanalysen gezeigt (MacCallum et al., 2007).

Nutzt man die aus Modellanpassungen gewonnen Itemparameter, so kann man für die Bestimmung der Personenparameter die inkorrekte Annahme der lokalen stochastischen Unabhängigkeit treffen. Dies führt im Allgemeinen zu konsistenten Schätzern (Clarke & Junker, 1991) und selbst bei einer kleinen Itemanzahl ist bei diesem Vorgehen nicht mit einer deutlich größeren Verzerrung als bei lokal stochastisch unabhängigen Items zu rechnen. Allerdings sind bei Verletzungen der lokalen stochastischen Unabhängigkeit individuelle Standardfehler unterschätzt (so dass die Reliabilität überschätzt wird), die jedoch unter Einsatz robuster Standardfehler korrigiert werden können (Ip, 2000; Ip & Chen, 2012). Alternativ stellt Haberman (2007) eine Methode zur Schätzung der Personenparameter(-verteilung) vor, die bei Modellverletzungen dieselbe bedingte Verteilung für die Item Responses wie unter der Modellgültigkeit besitzt (siehe auch Haberman & Sinharay, 2010, S. 213).

Brandt (2008, 2010) kritisiert die Annahme der bedingten Unkorreliertheit der Testletfaktoren im Testlet-Modell und schlägt als Modellalternative das sog. Rasch Subdimension Model vor, dass neben dem Trait θ die Testletfaktoren u_t enthält, die im Gegensatz zum Testlet-Modell korrelieren dürfen. Man kann allerdings zeigen, dass mit diesem Modell keine positive lokale Abhängigkeit modelliert wird, sondern vielmehr Testletfaktoren wie bei der Verwendung der Reliabilität ω_t als Bestandteil der wahren Varianz angesehen werden (man vgl. entsprechende Befunde in Brandt & Duckor, 2013)[12].

Testlets werden meistens in aufeinander folgenden Positionen in einem Test administriert. Demzufolge können positive lokale Abhängigkeiten innerhalb eines Testlets auch personenspezifische Ermüdungs- bzw. Positionseffekte innerhalb eines Tests widerspiegeln

[12] Das Rasch Subdimension Model ist ein reparametrisiertes mehrdimensionales Testlet-Modell, in dem jedes Testlet als eine Dimension definiert wird. Als Nebenbedingung muss die Summe der Testletfaktoren gleich Null sein, d.h. $\sum_t u_{nt} = 0$ für alle Personen n. Es wird $Cor(\theta, u_t) = 0$ für alle Testlets t angenommen. Interpretiert man die Anpassung des Rasch Subdimension Model als Kovarianzzerlegung einer tetrachorischen Korrelationsmatrix $\mathbf{\Sigma}$, so ergibt sich

$$\mathbf{\Sigma} = \underbrace{\sigma^2 \mathbf{1}\mathbf{1}' + \mathbf{\Phi}_{Testlets}}_{=:\mathbf{\Phi}} + \mathbf{\Theta} \tag{5.55}$$

Dabei ist $\mathbf{\Theta}$ eine Diagonalmatrix, die die lokale stochastische Unabhängigkeit abbildet und $\mathbf{\Phi}_{Testlets}$ die Kovarianzmatrix zwischen den Testletfaktoren. Wegen $\sum_t u_{nt} = 0$ folgt $\mathbf{1}'\mathbf{\Phi}_{Testlets}\mathbf{1} = 0$, so dass sich als Reliabilität gerade $\rho = I^2\sigma^2/(\mathbf{1}'\mathbf{\Sigma}\mathbf{1}) = (\mathbf{1}'\mathbf{\Phi}\mathbf{1})/(\mathbf{1}'\mathbf{\Sigma}\mathbf{1})$ ergibt. Demzufolge entspricht die Reliabilität aus dem Rasch Subdimension Model der Reliabilität aus einer Anpassung des mehrdimensionalen IRT-Modells mit Faktorkovarianzmatrix $\mathbf{\Phi}$.

Auf Basis dieser Überlegungen zu Brandts Rasch Subdimension Model kann man bei Einsatz des Testlet-Modells argumentieren, dass die Annahme von bedingt unkorrelierten Testleteffekten als ein spezielles Approximate Factor Model (vgl. Abschnitt 5.2.3) interpretiert werden kann, in dem man als Identifikationsannahme den Mittelwert zwischen den Testleteffekten als Null annimmt, jedoch die Residualkorrelationen zwischen den Testleteffekten nicht explizit schätzt, sondern diese im Testlet-Null auf Null setzt.

(Debeer & Janssen, 2013; Hartig & Buchholz, 2012; Robitzsch, 2009). Praktisch ist es schwer möglich, Testleteffekte von Positionseffekten zu trennen. Es wird häufig argumentiert, dass Positionseffekte formal eine Modellverletzung darstellen und deshalb explizit im Modell berücksichtigt werden müssten. Für die Definition des Traits θ würde dies zur Folge haben, dass die Fähigkeit nicht mehr als Testleistung in einer vorher fixierten Testzeit aufgefasst wird, sondern ein Referenzzeitpunkt im Test deklariert wird (zum Beispiel in der Mitte des Tests oder am Beginn des Tests), zu dem die wahre Varianz des Traits betrachtet wird. Alternativ könnten aus der Perspektive der Approximate Factor Models die Positionseffekte (analog zu den Testleteffekten) ignoriert werden. Unter der Identifikationsannahme, dass die mittlere Residualkovarianz Null ist, wird der Trait θ aus dieser Sichtweise als Fähigkeit definiert, die sich auf die gesamte Testzeit bezieht[13].

Lokale Abhängigkeiten spielen auch bei Längsschnittanalysen, in denen Items wiederholt einer Gruppe von Personen zur Messung eines Konstruktes vorgelegt werden, eine Rolle. Liegt ein Item z.B. zu zwei Zeitpunkten T1 und T2 vor, so wird man häufig eine positive lokale Abhängigkeit beobachten. Spezifiziert man diese positive lokale Abhängigkeit als Residualkorrelation in einem Messmodell, so verbessert sich im Allgemeinen der Modellfit (z.B. Cole, Ciesla & Steiger, 2007). Allerdings wird in der Literatur nur selten darauf hingewiesen, dass die Spezifikation einer Residualkorrelation auch die Definition des längsschnittlich zu erfassenden Traits θ ändert. Wenn die Abhängigkeiten ignoriert werden, beschreibt die Korrelation des Traits zu T1 und T2 die (minderungskorrigierte) Stabilität hinsichtlich derselben Items zu T1 und T2. Kontrolliert man dagegen Abhängigkeiten durch Spezifikation der Residualkorrelationen, so bezieht sich die Korrelation auf die Stabilität bezüglich der Beantwortung verschiedener Itemmengen zu T1 und T2. Auch wenn das Modell mit Residualkorrelationen besser passt, scheint es uns in vielen Anwendungen inhaltlich sinnvoller, die Stabilität bezüglich derselben Itemmenge zu erfassen.

Die Überlegungen unseres Beitrags werden treffend durch die folgende Aussage von Kane (2011) zusammengefasst

> Errors of measurement don't exist, until we create them. There is nothing about a test score per se that implies the existence of any errors of measurement.

Danach ist es für die Erfassung einer Kompetenz in einem Test mit Hilfe von Testlets zunächst unklar, wie der Fehler des Testwertes zu definieren ist. Die Auswahl eines psychometrischen Modells zur Definition des Fehlers hängt dann in jeder konkreten Anwendung davon ab, worauf der Testwert generalisiert werden soll.

[13] In Anlehung an das illustrative Datenbeispiel in §2 könnte man sich einen Test einer Testlänge von 30 Minuten vorstellen. Items in den ersten zehn Minuten betrachten wir dann als Testlet 1, in den zweiten zehn Minuten als Testlet 2 und in den dritten zehn Minuten als Testlet 3. Im eindimensionalen IRT-Modell M1 würde man die Abhängigkeiten ignorieren. Im Testlet-Modell M2 spezifiziert man neben einem Trait θ auch Testlet-Faktoren u_1, u_2 und u_3. Verwendet man das Reliabilitätsmaß ω_t, so bedeutet dies Fähigkeit (und damit assoziierte wahre Varianz) zu den drei Testabschnitten aufzufassen. Bei der Verwendung von ω_h werden Ermüdungseffekte (Testleteffekte u_t) als Fehlervarianz aufgefasst. Wie für Testlets in §2 argumentiert, ist die Bestimmung der „richtigen Reliabilität" nicht mittels eines Modellfits oder einer substanziellen Größe der Varianz der Ermüdungseffekte ableitbar.

Kapitel 6

Nichtignorierbare Konsequenzen des (partiellen) Ignorierens fehlender Item Responses im Large-Scale Assessment

In neuerer Literatur werden alternative Modellierungen fehlender Item Responses im Large-Scale Assessment vorgeschlagen. Prinzipiell argumentieren diese Ansätze auf Basis von Simulationen und testtheoretischer Überlegungen, dass fehlende Item Responses niemals als falsch zu bepunkten, sondern eher als ignorierbar zu behandeln seien (z.B. Pohl, Gräfe & Rose, 2014). Der vorliegende Beitrag zeigt auf, dass diese Argumentationen eingeschränkte Gültigkeit besitzen und illustriert anhand eines Ländervergleiches in PIRLS 2011, dass eine andere Bewertung der fehlenden Item Responses als die einer Falschantwort zu deutlichen Änderungen in Länderreihenfolgen führt und daher nichtignorierbare Konsequenzen hinsichtlich der Validität der Ergebnisse besitzt. Zusätzlich werden zwei alternative Item-Response-Modelle vorgeschlagen, mit denen verschiedene Annahmen für das Zustandekommen fehlender Werte bei den Item Responses beschrieben werden können.

6.1 Fehlende Item Responses in Large-Scale Assessments

In Large-Scale Assessments wie PIRLS, TIMSS, PISA oder den österreichischen Bildungsstandard-Erhebungen werden Kompetenzen von Schülerinnen und Schülern mit Hilfe von Testitems erfasst. Häufig geben Schülerinnen und Schüler allerdings keine Antwort bei bestimmten Items (fehlender *Item Response*; fehlende Itemantwort), so dass unklar ist, inwiefern sich eine Nichtbeantwortung eines Items auf die Bestimmung der Kompetenzwerte niederschlagen soll.

Während die Behandlung fehlender Daten im Rahmen statistischer Analysen in den Sozialwissenschaften mittlerweile verbreitet zu sein scheint (Graham, 2009; Lüdtke & Robitzsch, 2010), ist in jüngerer Literatur (Pohl et al., 2014; Rose, von Davier & Xu, 2010) Kritik an konventionellen Verfahren der Behandlung fehlender Item Responses in Item-Response-Modellen (IRT-Modelle) der Kompetenzmessung in Large-Scale Assessments zu finden. Typischerweise wird dabei die Missing-Behandlung zwischen dem Prozess der

Kalibrierung (Ermittlung von Itemparametern) und der Skalierung (Ermittlung von Kompetenzwerten) unterschieden.

In PIRLS/TIMSS und PISA werden bei der Kalibrierung ausschließlich fehlende Item Responses am Ende eines Testheftes (sog. *Not Reached* Items) weggelassen (d.h. ignoriert), um Schätzungen der Itemschwierigkeiten nicht zu „verzerren“. Für die Skalierung werden fehlende Item Responses typischerweise als falsch bewertet.

Pohl et al., 2014 und Rose, 2013 argumentieren jedoch, dass alle fehlenden Item Responses niemals (d.h. sowohl in Kalibrierung als Skalierung) als falsch zu kodieren seien und schlagen alternative Item-Response-Modelle zur Behandlung der fehlenden Itemantworten vor. Der folgende Beitrag geht dabei zunächst auf die in der Literatur angebrachten Kritikpunkte (Rose, 2013; Pohl et al., 2014) zur Behandlung fehlender Item Responses als Falschantwort und die vorgeschlagenen modellbasierten Alternativen ein. Wir argumentieren im Gegensatz zur o.g. Literatur, dass nur aus Gründen der Validität und nicht aus (scheinbaren) testtheoretischen Gründen eine bestimmte Methode für den Umgang mit fehlenden Item Responses präferiert werden sollte. Da im Allgemeinen Annahmen über den Datenausfall bei Item Responses empirisch nicht testbar sind, schlagen wir zwei alternative Item-Response-Modelle vor, in denen verschiedene Annahmen an den Ausfallprozess parametrisiert werden. Abschließend diskutieren wir mögliche psychometrische Konsequenzen im Large-Scale Assessment zur Erfassung von Kompetenzen bei unterschiedlichen Behandlungsweisen fehlender Itemantworten.

6.2 Eine Auseinandersetzung mit Kritikpunkten des „traditionellen“ Vorgehens bei fehlenden Item Responses

Für die Ermittlung von Kompetenzwerten werden im Large-Scale Assessment bei einer Skalierung typischerweise fehlende Item Responses als falsch bewertet. Konträr wird dazu in einer Reihe aktueller Publikationen (Rose, 2013; Pohl & Carstensen, 2012, 2013; Pohl et al., 2014) behauptet, dass fehlende Item Response niemals als falsch zu bewerten seien und es wird empfohlen, diese traditionelle „Ad-Hoc-Methode“ niemals im Large-Scale Assessment einzusetzen[1]. Im Folgenden gehen wir dabei auf die Begründungen der

[1]Im Folgenden möchten wir die Präferenz der Autoren für alternative Skalierungsmethoden zur konventionellen Bewertung als inkorrekt aufführen:

Rose et al. (2010, S. 44) resümieren: „*We conclude that treating missing data as wrong appears to be the least desirable way to account for responses MNAR in large-scale surveys. Model-based approaches seem to provide a more appropriate way to account for nonignorable missing data.*“
Pohl und Carstensen (2012, S. 8ff.) legitimieren das Skalierungsverfahren in NEPS wie folgt: „*[...] studies showed that ignoring missing responses, multiple imputation (Rubin, 1987), as well as model-based approaches (Glas & Pimentel, 2008; Holman & Glas, 2005; O'Muircheartaigh & Moustaki, 1999) result in unbiased parameter estimates, whereas treating missing responses as incorrect for either item or person parameter estimation results in biased parameter estimates. Pohl, Gräfe, and Hardt (2011), Gräfe (2012) as well as Pohl, Gräfe, and Rose (2012) compared the different approaches for treating missing responses in different domains and cohorts in NEPS and found indications that ignoring missing responses in the scaling model results in unbiased item and person parameter estimates. This closely resembles the results found in simulation studies (Rose, von Davier, & Xu, 2010). For scaling the competence data in NEPS,*

Autor/innen ein und gelangen zur Feststellung, dass diese auf Annahmen und Schlussfolgerungen beruhen, die untypisch in der Anwendung von Item-Response-Modellen sind und wir daher die in der o.g. Literatur befindlichen Empfehlungen für fragwürdig erachten.

---

all kinds of missing responses were thus ignored.“

Ausführlicher und unmissverständlich findet man diese Argumentation in Pohl und Carstensen (2013, S. 197ff.) „*The ignorability of the missing responses depends on the reason for responses to be missing. Whereas in most test designs missing responses due to not-administered items are missing by design and, therefore, missing completely at random or missing at random, omitted and not-reached items are usually nonignorable and often depend on the difficulty of the item and the ability of the person (Mislevy & Wu, 1988). If not treated correctly, nonignorable missing responses may lead to biased parameter estimates (Mislevy & Wu, 1996) and, thus, wrong conclusions about competence levels of persons as well as about relationships of competencies with other variables. [...] As Lord (1974) as well as Mislevy and Wu (1988, 1996) analytically derived, scoring missing responses as incorrect violates the model assumptions of IRT models. It induces a deterministic term and results in local item dependence. This has been corroborated in simulation studies and empirical analyses (e.g., Finch, 2008; Rose, von Davier, & Xu, 2010), which have shown that treating missing responses as incorrect results in biased estimates of item and person parameters. Ignoring omitted and not-reached items or using the model-based approaches leads to unbiased item and person parameter estimates. Model-based approaches do result in a higher reliability and allow us to investigate the ignorability of the missings. They are, however, more complex and require model assumptions (such as unidimensionality of the missing responses). [...] In empirical data, the approach of ignoring missing responses seems to be robust to violations of nonignorability. [...] Considerable differences in the person parameter estimates occur for those persons that have a large number of missing responses. The ability of these persons is, thus, heavily underestimated using the PISA approach as compared to ignoring the missing responses. The results found in empirical analyses using NEPS data are in line with results found in complete case simulation studies on the same data (Pohl et al., in press), thus supporting the superiority of the approach of ignoring missing responses as well as model-based approaches. Since the results show that the approach of ignoring missing responses is robust to violations of ignorability in these applications, it was decided to ignore missing responses in the scaling model of the NEPS competence data. In large-scale studies measuring competencies, model-based approaches can be used to investigate the amount of nonignorability and – if parameter estimates differ only slightly between model-based approaches and ignoring missing responses – to justify ignoring missing responses in the scaling model.“*

Etwas vorsichtiger wird in Pohl et al. (2014, S. 447ff.) argumentiert: „*The IRT model seemed to be quite robust to violations of the ignorability assumption for missing responses in empirical settings like the one we investigated here. This was corroborated in the complete case simulation where we simulated different missing mechanisms. This is in line with findings from Hoshino (2005), who concluded that the observed item responses (or even just the latent variables in the measurement model) are sufficient to account for missing responses in competence tests. [...] However, ignoring missing responses in the estimation may encourage test takers to omit items as a test taking strategy. This approach may therefore be inappropriate for high-stakes assessments, and model-based approaches may be a good alternative. [...] As could be shown, IRT-based item and person parameter estimates turned out to be fairly robust to violation of the MAR assumption (Rose, 2013). However, model based approaches do not only reduce bias but can increase accuracy of parameter estimates. For example, even if the MAR assumption holds true, missingness can be correlated with the latent ability. Using model-based approaches reduces the shrinkage effect and standard errors of EAP and maximum a posteriori person parameter estimates (Rose, 2013). Recently, Rose (2013), Rose, von Davier, and Nagengast (2013), and Rose and von Davier (2013) introduced further alternative models and stepwise procedures to handle nonignorable missing data due to omitted and not reached items. Similar to the approaches discussed in this article, these methods allow us to adjust for nonignorable missing data as well as increase the accuracy of item and person parameter estimates.“*

Eine eindeutige Ablehnung der Bewertung fehlender Item Responses als falsch findet man in Rose (2013,

6.2.1 Aleatorische und epistemische Unsicherheit

Nützlich für die folgenden Überlegungen erweist sich dabei die von Denoeux (2011) vorgeschlagene Unterscheidung bei der Quantifikation der Unsicherheit in Daten (den Item Responses) und damit korrespondierenden Modellierungen, die in der Abgrenzung der Konzepte der „Probability“ (Wahrscheinlichkeit) und „Possibility“ (Möglichkeit) bedeutsam sind. Durch Stichprobenziehung oder Modellierung existierender Daten (Datensatz mit allen Item Responses) entstehende Unsicherheit wird dabei als *aleatorische Unsicherheit* bezeichnet. Unsicherheit (bzw. Unschärfe in der Terminologie von Fuzzy-Daten) für jedes einzelne Datum (jede einzelne Itemantwort) wird als *epistemische Unsicherheit* bezeichnet. Denoeux (2011) argumentiert, dass man aleatorische Unsicherheit mittels Wahrscheinlichkeiten modellieren solle, was im Falle des vorliegenden Kompetenztests bedeutet, ein IRT-Modell für fest vorgegebene Item Responses anzupassen. Die Beobachtung der einzelnen Daten (der einzelnen Item Responses jeder Schülerin oder jedes Schülers) erfolgt dabei mit dem Konzept epistemischer Unsicherheit, das in Anwendungen dann häufig in unscharfen Beobachtungen (sog. Fuzzy-Daten) mündet (Denoeux, 2011, 2013). Bei fehlenden Item Responses ist demzufolge unklar, welche Bepunktung das Fehlen eines Item Response nach sich ziehen soll. Es ist aber in dieser Konzeption offensichtlich, dass Annahmen an die einzelnen Daten (den einzelnen Item Responses) strikt von Annahmen in einem Wahrscheinlichkeitsmodell (ein gemeinsames Modell für alle Item Responses) zu unterscheiden sind. Zur Illustration dieser begrifflichen Unterscheidung nehmen wir an, dass man die Verteilung der Laufzeiten in einem 50-Meter-Lauf für eine definierte Population österreichischer Schüler/innen der 4. Schulstufe erfassen möge. Dazu wird eine Stichprobe von Schüler/innen gezogen. Die mit dieser Stichprobenziehung verbundene Unsicherheit auf die Population wird mit Wahrscheinlichkeiten im Kontext der aleatorischen Unsicherheit ausgedrückt. Jede Messung der Laufzeit eines Schülers ist aber niemals exakt, sondern immer nur unscharf (z.B. liegt die gemessene Laufzeit aufgrund der Messunsicherheit im Intervall zwischen 8.75 und 8.85 Sekunden, das ein Fuzzy-Datum darstellt). Diese Unsicherheit entspricht dem Konzept epistemischer Unsicherheit.

6.2.2 Testtheoretische „Begründungen“

Die Kritik der Falschbewertung fehlender Item Responses fußt dabei auf simulationsbasierten und testtheoretischen Begründungen (Rose, 2013), auf die wir in weiterer Folge jeweils näher eingehen.

Die meisten Simulationsstudien versuchen zu vermitteln, dass eine Falschbewertung von fehlenden Item Responses zu „verzerrten“ Itemparametern und Personenparametern führt. Dabei wird ein fehlender Item Response meist durch eine Abhängigkeit von einer weiteren latenten Variable oder Personenkovariaten (nicht aber in Abhängigkeit des Items selbst) simuliert. Dann ist die berichtete Verzerrung bei einer Bewertung als falsch

S. 298ff.): „*Based on the results of this work and in line with previous research, different recommendations for applied researchers in the field of educational and psychological measurement can be derived. At first it is strongly recommended never to use ad hoc methods such as IAS [incorrect answer substitution; Bewertung als falsch] or PCS [partial correct scoring] to handle item nonresponses. Simply to ignore missing data in IRT models seems to be less harmful than using such ad hoc methods.*“

in Parametern folgerichtig, denn das der Simulation zugrunde liegende datengenerierende Modell stimmt nicht mit dem Analysemodell (Missings sind falsche Itemantworten) überein. Nimmt man jedoch umgekehrt an, dass ein IRT-Modell für die Item Responses passt und es werden bei allen Items fehlende Werte generiert, für die das Item falsch gelöst wurde, dann führt natürlich jede zur Falschbewertung alternative Missingbehandlung zu verzerrten Parametern (siehe Rohwer, 2013). Ob eine simulationsbasierte Argumentation gegen eine Falschbewertung von fehlenden Item Responses zutreffend ist, hängt demzufolge von der Plausibilität des datengenerierenden Modells für den Ausfallprozess ab. Wir argumentieren, dass Ansätze, die den Ausfall auf einem Item unabhängig von der unbekannten Itemantwort selbst modellieren, unplausibel sind. Somit helfen Simulationsstudien bei der Frage der Begründung einer „richtigen Bewertung“ fehlender Item Responses nicht weiter.

In der testtheoretischen Begründung ist der Kern der Kritik, dass eine Falschbewertung fehlende Items „deterministisch“ behandeln würde und demzufolge modellimplizierte Wahrscheinlichkeiten des IRT-Modells ungültig seien (Rose, 2013; Pohl et al., 2014). Bezeichnen wir mit R_{pi} die Dummyvariable, dass für Person p das Item i beobachtet wird, so wird bei einer Falschbewertung der Item Response Y_{pi} auf Null (d.h. falsch) gesetzt, also $P(Y_{pi} = 1|R_{pi} = 0) = 0$ unabhängig von der Personenfähigkeit θ_p. Dadurch entstünde ein Widerspruch, denn über die Vorhersage im IRT-Modell (z.B. dem Rasch-Modell) würde man eine Wahrscheinlichkeit $P(Y_{pi} = 1|\theta_p) > 0$ annehmen, d.h. die Datenbehandlung sei deterministisch und das IRT-Modell sei probabilistisch (Rose, 2013), weshalb Modellannahmen im IRT-Modell verletzt seien[2]. Formal ist der Widerspruch mit den beiden obigen Überlegungen verschiedener Konzepte der Unsicherheit nach Denoeux (2011) auflösbar, denn in der Argumentation von Rose (2013) und Pohl et al. (2014) erfolgt eine Konfundierung von epistemischer und aleatorischer Unsicherheit, so dass man damit gar nicht von einer Gleichsetzung der verschiedenen Wahrscheinlichkeiten ausgehen sollte. Betten wir unsere Überlegungen nun stärker in Begrifflichkeiten der Psychometrie ein, so fällt auf, dass die Argumentation der Ablehnung als Falschbewertung offenbar auf einer intraindividuellen Interpretation der Wahrscheinlichkeiten im IRT-Modell im Sinne der *Stochastic Subject Perspektive* (d.h. Wahrscheinlichkeiten des IRT-Modells können für jede Kombination einer Person n mit einem Item i interpretiert werden) beruht (siehe Holland, 1990a; für eine auf der Stochastic Subject basierende Testtheorie siehe Steyer & Eid, 2001), die in Anwendungen des Large-Scale Assessments allenfalls historische und geringe praktische

[2]Die Begründung von Rose (2013) beruht auf zwei kritischen Annahmen. Erstens wird angenommen, dass die Fehlervariablen intraindividuell definiert werden. Damit ist der True Score eines Items nur durch die latente Variable im Messmodell definiert. Wir schließen uns dem Standpunkt von McDonald (2011, S. 516) zur intraindividuellen Perspektive an: “Such a sample space does not seem necessary and certainly has no counterpart in applications.“ Zweitens nimmt Rose (2013) an, dass Messinvarianz „gilt“. Messinvarianz bedeutet dabei, dass die Verteilung eines Items i gegeben die latente Variable θ invariant für alle Werte einer Kovariaten Z ist, d.h. $P(Y_i|\theta, Z) = P(Y_i|\theta)$. Rose setzt dann die Response-Indikatoren R_i anstelle von Z ein, so dass folgt $P(Y_i|R_i = 1, \theta) = P(Y_i|R_i = 0, \theta) = P(Y_i|\theta)$. Dies zeigt aber die Zirkularität seiner Argumentation. Die latente Variable θ im Messmodell ist demzufolge nicht nur über die unbeobachtete Verteilung der Items $P(\boldsymbol{Y})$ definiert, sondern bezieht bereits die Response-Indikatoren R_i ein. Auch das Konzept der Messinvarianz ist damit *nicht* nur auf Basis der unbeobachteten Verteilung $P(\boldsymbol{Y})$ definiert. Wir würden sogar generell hinterfragen, ob Invarianz ein relevantes Konzept für Messmodelle mit latenten Variablen darstellt (siehe auch Abschnitt 7.3).

Bedeutung besitzt (Wainer, 2010b). Stellt man eine *Random Sampling* Perspektive (Holland, 1990a; Molenaar, 1995) an, so interpretiert man Wahrscheinlichkeiten als Ergebnis einer Stichprobenziehung von Personen und führt Itemparameter sowie eine Verteilung für Personenfähigkeiten θ als Repräsentation einer hochdimensionalen Kontingenztafel für multivariate diskrete Beobachtungen $\boldsymbol{Y}$ ein. Dann ergibt sich formal als statistisches Modell eine Repräsentation als ein Integral

$$P(\boldsymbol{Y} = (y_1, \ldots, y_I)) = \int_\theta \prod_{i=1}^{I} P(Y_i = y_i|\theta) f(\theta)\, \mathrm{d}\theta \tag{6.1}$$

In dieser Schreibweise wird deutlich, dass in IRT-Modellen eine Verteilung für Personen spezifiziert wird und nicht einzelne Individuen repräsentiert werden (siehe auch Rohwer, 2013).

Wenn man jedoch die intraindividuelle Perspektive im Sinne der *Stochastic Subject* Perspektive annimmt, so beruht die Argumentation auf der Behauptung, dass bei einem Scoring der Missings als falsch die Wahrscheinlichkeit für eine korrekte Lösung eines Items deterministisch sei, denn die Bewertung als falsch führe zu einer falschen Antwort mit Wahrscheinlichkeit 1. Die Bewertung einer Antwort als falsch steht aber in keinem Zusammenhang zur probabilistischen Modellierung der Itemantworten einer Person. Denn es könnten auch die nichtfehlenden Items durch einen wie auch immer gearteten deterministischen Prozess zustande gekommen sein, was nicht der probabilistischen Modellierung im statistischen Modell für alle Item Responses widerspricht. Man muss daher zwischen einem „realen Antwortprozess“ – der deterministisch oder probabilistisch erfolgen kann, aber sicher in den seltensten Fällen der vorgegebenen probabilistischen Spezifikation des IRT-Modells folgt – und der statistischen Modellannahme zur Definition einer Fähigkeit im Sinne einer Skalenkonstruktion unterscheiden (Rohwer, 2013).

Auch statistische Gründe im Sinne der Stochastic Subject Perspective sprechen entkräften die Kritik an einer Falschbewertung der fehlenden Item Responses. Bei der Modellierung intraindividueller Verteilungen in einem IRT-Modell für eine Person werden dabei Itemparameter als feste und bekannte Parameter vorausgesetzt und die Personenfähigkeit θ_p ist als einer der Person zugeordneter fester Effekt (*fixed effect*) zu interpretieren. Die Annahme der lokalen stochastischen Unabhängigkeit bezieht sich dann auf eine bedingte Unabhängigkeit der Itemantworten für eine feste Person p anhand der vorgegebenen Itemmenge und ist damit empirisch nicht widerlegbar, sondern eine Setzung zur Identifikation der Fähigkeit θ_p dieser Person. D.h. die Personenfähigkeit θ_p ist erst durch die Spezifikation der Likelihood und der Itemantworten $\boldsymbol{Y}$ definiert. Im Rasch-Modell ist bei einer festen Person p und Item i die Wahrscheinlichkeit einer korrekten Itemantwort durch $P(Y_{pi} = 1|\theta_p) = \Psi(\theta_p - b_i)$ (mit der logistischen Funktion Ψ) gegeben, die in der Likelihood-Schätzung für den Parameter θ_p verwendet wird. Ob eine einzelne Beobachtung dann in diesem Modell „passt“, ist empirisch nicht entscheidbar, so dass eine Falschbewertung der fehlenden Item Responses nicht per se das IRT-Modell verletzen kann.

Zusammenfassend sind die in der Literatur (Rose, 2013; Pohl et al., 2014) auffindbaren simulationsbasierten und testtheoretischen Begründungen, die gegen die Bewertung der fehlenden Antworten als falsch angebracht werden, aus unserer Sicht für typische Anwendungen im Large-Scale Assessment nicht zutreffend. Wir merken jedoch an, dass aus

Gründen der Validität durchaus andere Bewertungen als falsch in Anwendungen adäquat sein könnten. Zentral in der Argumentation der Kritik der Autor/innen ist, dass aus dem Antwortverhalten von Schüler/innen in einer querschnittlichen Messung (interindividuelle Perspektive) fälschlicherweise auf ein Antwortverhalten eines einzelnen Schülers (intraindividuelle Perspektive) rückgeschlossen wird (siehe Molenaar, 2004).

6.3 Modellbasierte Behandlung fehlender Item Responses

In diesem Abschnitt werden verschiedene modellbasierte Verfahren zur Behandlung fehlender Item Responses diskutiert, wobei wir im Folgenden nur IRT-Modelle für dichotome Daten in der Familie der Rasch-Modelle (Fischer & Molenaar, 1995) betrachten. Allerdings sind unsere Überlegungen auch für allgemeinere Modellklassen wie 2PL-Modelle oder IRT-Modelle für polytome Daten (Yen & Fitzpatrick, 2006) gültig.

Unter den modellbasierten Verfahren werden häufig IRT-Modelle für ignorierbare und nichtignorierbare Item Responses unterschieden (Holman & Glas, 2005). Bei einer Ignorierung fehlender Item Responses werden Items für Schüler/innen aus der Likelihood-Funktion in der Schätzung weggelassen (d.h. ignoriert), für die das jeweilige Item fehlend ist. Ignoriert man die fehlenden Item Responses in der Schätzung, so kann man zeigen, dass fehlende Item Responses unter dieser Annahme mit einer Wahrscheinlichkeit von $P_{pi}(\theta_{p,M})$ als richtig imputiert (d.h. bepunktet) werden, wenn $\theta_{p,M}$ die Fähigkeit von Person p bezeichnet, die ausschließlich mit den nichtfehlenden Items ermittelt wurde. D.h. praktisch, dass das IRT-Modell und die beobachteten Item Responses genutzt werden, um die fehlenden Items zu ersetzen[3]. Da die Wahrscheinlichkeit $P_{pi}(\theta_{p,M})$ immer größer als Null ist, führt die Ignorierung von Item Responses bei fest gehaltenen Itemparametern immer zu höheren Personenfähigkeiten als eine Bewertung als falsch. Das Ignorieren fehlender Item Responses bedeutet nicht, dass der Ausfall unabhängig von der Fähigkeit ist (d.h. der Ausfall ist nicht missing completely at random, MCAR; siehe Lüdtke & Robitzsch, 2010). Er ist dadurch charakterisiert, dass die gesamte Information über die Fähigkeit θ_p bereits aus den beobachteten Item Responses rekonstruierbar ist, d.h. der Ausfall ist missing at random (MAR). D.h. unter dieser Annahme kann der Fall eintreten, dass der Anteil fehlender Item Responses bei leistungsschwächeren Schüler/innen größer als bei leistungsstärkeren Schüler/innen ist.

Die Behandlung fehlender Item Responses als falsch kann dabei als ein Extremum, die Ignorierung als ein anderes Extremum angesehen werden (Rost, 2004, S. 324 ff.). Im ersteren Fall besteht die Gefahr einer Unterschätzung der Personenfähigkeit, im zweiten Fall gegebenenfalls eine Überschätzung der Personenfähigkeit.

Alternativ zur Ignorierung fehlender Item Responses wurden mehrdimensionale IRT-

[3]Unter der Annahme der Ignorierbarkeit der fehlenden Item Responses sagt man $P(Y_i = 1|\theta_p)$ für den fehlenden Response Y_i vorher. Daher wird das Item mit der Wahrscheinlichkeit $P_{pi}(\theta_{p,M})$ imputiert. Die latente Variable θ_p muss daher aus den beobachteten Responses rekonstruiert werden. Im Rasch-Modell können wir $P(Y_{pi} = 1) \propto \exp(1 \cdot (\theta_p - b_i))$ sowie $P(Y_{pi} = 0) \propto \exp(0 \cdot (\theta_p - b_i))$ schreiben. Damit gilt $\prod_i P(Y_{pi} = 1) \propto \exp(\sum_i y_{pi}(\theta_p - b_i))$ und es folgt das Scoring der fehlenden Item Responses mit $P_{pi}(\theta_{p,M})$

Modelle für nichtignorierbare Item Responses vorgeschlagen (Holman & Glas, 2005; Rose et al., 2010). Das zweidimensionale Modell von Holman und Glas (2005) führt neben der latenten Fähigkeit θ eine latente individuelle Response Propensity (Response-Tendenz) ξ ein, die dem Ausfall der Item Responses zugrunde liegt. Für die Responsevariablen R_{pi} nimmt man dabei ebenso ein Rasch-Modell an:

$$P(R_{pi} = 1|\theta_p, \xi_p) = \Psi(\xi_p - \beta_i) \tag{6.2}$$

Die Wahrscheinlichkeit, eine Itemantwort zu produzieren, hängt demnach von der Response-Tendenz ξ_p und von der „Schwierigkeit" eines nichtfehlenden Responses β_i ab. Ob ein Item fehlt oder nicht, hängt allerdings nicht vom Item Response selbst ab. Das zweidimensionale Modell schätzt damit eine bivariate Verteilung von (θ, ξ). Wir bemerken, dass für jede Person keine Maximum-Likelihood-Schätzung für die Fähigkeit θ_p, sondern nur eine EAP-Schätzung (Yen & Fitzpatrick, 2006) existiert. Als Konsequenz ist dann nur unter Vorgabe einer Korrelation von Fähigkeit θ und Response-Tendenz ξ eine Schätzung der Fähigkeit möglich.

Lässt sich dann die Response-Tendenz ξ als $\xi = \rho\theta + \varepsilon$ schreiben, so kann man zeigen, dass fehlende Item Responses näherungsweise eine Bepunktung von $P_{pi}(\theta_{p,M}) - \rho$ erfahren (siehe Bertoli-Barsotti & Punzo, 2013). Im Unterschied zum Modell mit ignorierbaren Item Responses werden Fähigkeitsschätzungen demzufolge in Abhängigkeit der Regression der Response-Tendenz auf die Fähigkeit um die Konstante ρ adjustiert. Nur wenn Response-Tendenz und Fähigkeit unkorreliert sind ($\rho = 0$), führen beide Modelle zu denselben Personenparameterschätzungen.[4]

Köhler, Pohl und Carstensen (2015) untersuchen nichtnormalverteilte bivariate Verteilungen für den Vektor (θ, ξ) der Fähigkeit und der Response-Tendenz. Dabei wird die bivariate Verteilung diskretisiert und einer log-linear geglättet (Xu & von Davier, 2008). Unter der Nichtnormalverteilungsannahme besitzt die Response-Tendenz nach Köhler et al. (2015) im Gegensatz zur Normalverteilungsannahme einen nichtvernachlässigbaren Einfluss auf die Fähigkeitsschätzung.

Als Approximation des zweidimensionalen Modells schlagen Rose et al. (2010; siehe auch Pohl et al., 2014) ein eindimensionales IRT-Modell (bei Ignorierung fehlender Item Responses) für die Fähigkeit mit einem latenten Hintergrundmodell vor, in dem der Anteil fehlender Items $\tilde{\xi}$ als manifeste Kovariate verwendet wird. Diese Kovariate soll dabei näherungsweise die Rolle der Response Propensity ξ einnehmen. Die Posteriorverteilung von θ lässt sich dann schreiben gemäß (vgl. Adams & Wu, 2007)

$$P(\theta|\boldsymbol{Y}, \tilde{\xi}) \propto P(\boldsymbol{Y}|\theta)P(\theta|\tilde{\xi}) \tag{6.3}$$

Daran erkennt man, dass sich die Posterior aus der individuellen Likelihood $P(\boldsymbol{Y}|\theta)$ und der empirischen Priorverteilung $P(\theta|\tilde{\xi})$ zusammensetzt. Bei längeren Tests mit vielen

[4] Für beobachtete und korrekte Item Responses schreibt sich der Zähler der gemeinsamen Verteilung als $P(Y_{pi} = 1, R_{pi} = 1) \propto \exp[1 \cdot (\theta_p - b_i) + 1 \cdot (\xi_p - \beta_i)] \propto \exp[(1 + \rho)\theta_p + \epsilon_p]$. Für inkorrekte Item Responses gilt $P(Y_{pi} = 0, R_{pi} = 1) \propto \exp[0 \cdot (\theta_p - b_i) + 1 \cdot (\xi_p - \beta_i)] \propto \exp[(0 + \rho)\theta_p + \epsilon_p]$. Daraus folgt das Scoring von $1 + \rho$ bzw. von ρ für korrekte bzw. inkorrekte beobachtete Item Responses. Für unbeobachtete Responses Y_{pi} ist $P(Y_{pi} = y_{pi}, R_{pi} = 1) \propto \exp[y_{pi} \cdot (\theta_p - b_i) + 0 \cdot (\xi_p - \beta_i)] \propto \exp[y_{pi}\theta_p]$. Da y_{pi} mit Wahrscheinlichkeit $P_{pi}(\theta_{p,M})$ mit 1 imputiert wird, erfolgt für fehlende Item Responses ein Scoring von $P_{pi}(\theta_{p,M})$. Subtrahiert man von den gerade abgeleiteten Scoringvorschriften jeweils ρ, so folgt die Behauptung.

Items dominiert bei der Schätzung der Personenfähigkeit die Likelihood der nichtfehlenden Item Responses im Vergleich zur Priorverteilung des latenten Hintergrundmodells, so dass man für Tests mit hinreichend vielen Items näherungsweise das Vorgehen der Ignorierbarkeit fehlender Item Responses erhält.

Weitere modellbasierte Verfahren verallgemeinern das bivariate IRT-Modell von Holman und Glas (siehe Rose, 2013; Bertoli-Barsotti & Punzo, 2013) oder setzen Mischverteilungsansätze ein (Bacci & Bartolucci, 2013; Pietsch, 2011). Wir beziehen diese Modelle allerdings aus Gründen der Übersichtlichkeit nicht in die vergleichenden Analysen dieses Kapitels ein.

6.4 Zwei alternative Item-Response-Modelle für nichtignorierbare Item Responses: Ansätze für eine Sensitivitätsanalyse

In diesem Abschnitt sollen zwei alternative modellbasierte Ansätze diskutiert werden, die die fehlenden Item Responses als nichtignorierbar behandeln. Im ersten Modell soll der Ansatz der teilrichtigen Bewertung von Lord (1974) angewendet werden. Ein zweites IRT-Modell erweitert das zweidimensionale IRT-Modell von Holman und Glas (2005) um eine mögliche Abhängigkeit des Ausfallens eines Items vom unbekannten Item Response selbst, so dass sich die Bewertung einer fehlenden Itemantwort als falsch als ein Spezialfall ergibt. Die beiden vorgeschlagenen Modelle können dabei als eine so genannte Sensitivitätsanalyse angesehen werden, bei der unter einer Variation verschiedener Annahmen an den Ausfallprozess die erhaltenen Ergebnisse studiert werden (z.B. Resseguier et al., 2011).

6.4.1 Pseudo-Likelihood-Ansatz für partielles Scoring der Item Responses

Im Pseudo-Likelihood-Ansatz nach Lord (1974) können Item Responses als partiell korrekt (teilrichtige Bepunktung) bewertet werden. Die Pseudo-Likelihood-Funktion L_p (genauer: die Pseudo-Log-Likelihood) für Person p ist definiert als

$$\log L_p = \sum_{i=1}^{I} [w_{pi} \log P_{pi} + (1 - w_{pi}) \log(1 - P_{pi})] \tag{6.4}$$

Dabei ist P_{pi} die Wahrscheinlichkeit der korrekten Beantwortung von Item i durch Person p und w_{pi} ist der Punktwert (Score) von Person p auf Item i. Liegen keine fehlenden Item Responses vor, dann sind die Scores w_{pi} entweder gleich 1 (bei einer richtigen Antwort) oder gleich 0 (bei einer falschen Antwort). Lord (1974) argumentiert, dass bei fehlenden Item Responses bei Multiple Choice Items mit M Antwortalternativen die Scores w_{pi} gleich einer Ratewahrscheinlichkeit von $1/M$ gesetzt werden können und die entstehende Likelihood (6.4) optimiert wird. Würde man die fehlenden Item Responses ignorieren, dann bedeutet dies – wie oben argumentiert – fehlende Items mit $w_{pi} = P_{pi}(\theta_{p,M})$ zu bepunkten, wobei $\theta_{p,M}$ die Fähigkeitsschätzung für Person p auf Basis der nichtfehlenden

Item Responses ist. Bei einer Behandlung der fehlenden Item Responses als falsch wählt man die Scores $w_{pi} = 0$. Diese beiden Fälle sollen als „Extrema“ der Behandlung fehlender Item Responses aufgefasst werden (siehe Rost, 2004, S. 324 ff.) und „dazwischen liegende“ Annahmen durch einen *Sensitivitätsparameter* $\rho = 0$ (Behandlung der Missings als falsch) und $\rho = 1$ (Behandlung der Missings als ignorierbar) im Rahmen einer Sensitivitätsanalyse studiert werden. Wir setzen in diesem Ansatz als Bepunktung von Item i für Person p der Score $w_{pi} = \rho P_{pi}(\theta_{p,M})$ und ermitteln die Ergebnisse der Skalierung bei Variation des Parameters ρ. Diese Technik der Sensitivitätsanalysen ist bei der Imputation fehlender Daten für nichtignorierbare Missings verbreitet (van Buuren, 2012; Resseguier et al., 2011). Dabei ist zu bemerken, dass der Parameter ρ selbst nicht schätzbar ist, sondern für die Schätzung vorgegeben werden muss. Dies wird auch dadurch gestützt, dass die Pseudo-Likelihood-Funktion in Abhängigkeit von ρ nur ein Maximum bei $\rho = 0$ oder $\rho = 1$ annehmen kann[5].

Die Likelihood-Funktion (6.4) lässt sich schreiben als

$$L_p = \prod_{i=1}^{I} L_{pi} = \prod_{i=1}^{I} \left\{ P_{pi}^{w_{pi}} (1 - P_{pi})^{1-w_{pi}} \right\} \tag{6.6}$$

Demzufolge hat die Likelihood für Person p auf Item i den Beitrag $L_{pi} = P_{pi}^{w_{pi}}(1 - P_{pi})^{1-w_{pi}}$. Die nicht beobachteten Items lassen sich demzufolge als unscharfe Daten (Fuzzy-Daten) mit Werten 0 und 1 und zugehöriger *membership function* $m_{pi}(0) = 1 - w_{pi}$ bzw. $m_{pi}(1) = w_{pi}$ auffassen (Denoeux, 2013). Der multiplikative Term L_{pi} wird auch als *partial membership* bezeichnet (siehe Gruhl & Erosheva, 2015, S. 21). Alternativ wurden für unscharfe Daten auch ein additiver Likelihood-Term gemäß $L_{pi} = w_{pi}P_{pi} + (1 - w_{pi})(1 - P_{pi})$ vorgeschlagen (vgl. Denoeux, 2013), der auch als *mixed membership* bezeichnet wird (Gruhl & Erosheva, 2015, S. 20).

Modellschätzung

Für die Schätzung im Pseudo-Likelihood-Ansatz (6.4) ist die Definition der Gewichte w_{pi} notwendig. Diese individuellen und itemspezifischen Gewichte müssen in einem ersten Skalierungsschritt gewonnen werden. Dabei werden Itemschwierigkeiten unter einer Annahme der Behandlung der Missings gewonnen (Behandlung als falsch oder als ignorierbar) und danach individuelle Personenschätzer (z.B. weighted likelihood estimates, siehe Warm, 1989) auf Basis der beobachteten Daten und fixierten Itemschwierigkeiten ermittelt. Alternativ könnten auch Plausible Values aus der individuellen Posteriorverteilung gezogen werden (Mislevy, 1991). Der Pseudo-Likelihood-Ansatz (6.4) mit fixierten Gewichten w_{pi} kann dabei mit dem üblichen EM-Algorithmus (siehe z.B. von Davier & Sinharay, 2014)

[5] Wir schreiben $w_{pi} = \rho P_{pi}(\theta_{p,M}) = \rho\alpha_{pi}$. Dann gilt für die Log-Likelihood $f = \log L_p$ als Funktion von ρ die Beziehung

$$f(\rho) = \log L_p(\rho) = \sum_i \log(1 - P_{pi}) + \rho \sum_i \alpha_{pi} \left\{ \log P_{pi} - \log(1 - P_{pi}) \right\} \tag{6.5}$$

Als lineare Funktion ρ kann damit die Likelihood nur für $\rho = 0$ oder $\rho = 1$ ein Maximum besitzen.

geschätzt werden, wobei der M-Step gegenüber dem gewöhnlichen Rasch-Modell unverändert bleibt, die Azuswertung der individuellen Likelihood und der expected counts nun allerdings die Pseudo-Likelihood anstelle der gewöhnlichen Likelihood verwendet.

6.4.2 Modellierung des Ausfallprozesses der Item Responses

Im zweiten Modell werden die fehlenden Item Responses ähnlich dem zweidimensionalen Modell von Holman und Glas (2005; siehe auch Rose et al., 2010) modelliert. Schreiben wir wiederum Ψ als Abkürzung für die logistische Funktion, so gilt unter der Annahme des Rasch-Modells (Yen & Fitzpatrick, 2006) folgende Gleichung für die Wahrscheinlichkeit eines korrekten Item Responses

$$P(Y_i = 1|\theta) = \Psi(\theta - b_i) \tag{6.7}$$

Die Modellgleichung für einen fehlenden Wert des Items i mit Hilfe der Response Propensity (Ausfalltendenz) ξ möge nun allerdings auch vom unbekannten Item Response Y_i selbst abhängen (siehe Mislevy & Wu, 1996). Dafür definieren wir

$$P(R_i = 1|Y_i = k, \xi) = \Psi(\xi - \beta_i - k\delta) \quad \text{mit } k = 0, 1 \tag{6.8}$$

Die Wahrscheinlichkeit eines nichtfehlenden Item Response (d.h. $R_i = 1$) in Abhängigkeit des Item Response Y_i entsteht dabei unter der Annahme $\delta \neq 0$. Wählt man $\delta = 0$, so gelangt man zum IRT-Modell für nichtignorierbare Item Responses von Holman und Glas (2005). Wiederum können im Rahmen einer Sensitivitätsanalyse Skalierungsergebnisse in Abhängigkeit vorgegebener Parameterwerte für δ studiert werden. Für sehr kleine Werte von δ (also z.B. $\delta = -10$) erhält man $P(R_i = 1|Y_i = 1, \xi) = 1$, also $P(R_i = 0|Y_i = 1, \xi) = 0$. D.h. alle Schüler/innen, die die korrekte Itemantwort wissen, geben mit einer Wahrscheinlichkeit von 1 auch einen korrekten Item Response und lassen daher dieses Item nicht aus. Daraus folgt allerdings nach dem Satz von Bayes die Beziehung $P(Y_i = 1|R_i = 0; \theta, \xi) = 0$, d.h. fehlende Item Responses werden mit einer Wahrscheinlichkeit von 1 als falsch bewertet.

Wie im Pseudo-Likelihood-Ansatz modellieren wir demzufolge auch mit diesem Modell die Extrema der Ignorierbarkeit von Missings ($\delta = 0$) und der Behandlung von Missings als falsch ($\delta = -10$).

Das vorgeschlagene Item-Response-Modell schätzt Itemschwierigkeiten b_i, Itemparameter β_i für die Response-Tendenz sowie die bivariate Verteilung von (θ, ξ). Dabei werden in der Ermittlung der Likelihood für Item i die Wahrscheinlichkeiten dreier disjunkter Ereignisse $P(R_i = 1, Y_i = 0|\theta, \xi)$, $P(R_i = 1, Y_i = 1|\theta, \xi)$ und $P(R_i = 0|\theta, \xi)$ verwendet. Die Wahrscheinlichkeit eines fehlenden Item Responses $P(R_i = 0|\theta, \xi)$ ist dabei durch die totale Wahrscheinlichkeit

$$P(R_i = 0|\theta, \xi) = P(R_i = 0|Y_i = 0; \xi)P(Y_i = 0|\theta) + P(R_i = 0|Y_i = 1; \xi)P(Y_i = 1|\theta) \tag{6.9}$$

gegeben, die man aus den Itemparametern berechnen kann. Fehlende Item Responses Y_i werden dabei praktisch in Analogie zu fehlenden Items in Large-Scale Assessments mit der Marginal Maximum Likelihood (MML) Schätzmethode ausintegriert (von Davier & Sinharay, 2014; siehe auch Mislevy & Wu, 1996; Hanson, 2000).

Modellschätzung

Das vorgeschlagene IRT-Modell beruht auf der Auswertung von Response-Wahrscheinlichkeiten für die latenten Variablen θ und ξ. Für beobachtete korrekte Item Responses gilt mit (6.7) und (6.8)

$$P(Y_i = 1, R_i = 1|\theta, \xi) = P(R_i = 1|Y_i = 1, \xi) \cdot P(Y_i = 1|\theta) = \Psi(\xi - \beta_i - \delta) \cdot \Psi(\theta - b_i) \quad (6.10)$$

Für beobachtete inkorrekte Item Responses gilt

$$P(Y_i = 0, R_i = 1|\theta, \xi) = P(R_i = 1|Y_i = 0, \xi) \cdot P(Y_i = 0|\theta) = \Psi(\xi - \beta_i) \cdot \Psi(-\theta + b_i) \quad (6.11)$$

Für fehlende Item Responses setzt man mit (6.9) an

$$P(R_i = 0|\theta, \xi) = \Psi(-\xi + \beta_i) \cdot \Psi(-\theta + b_i) + \Psi(-\xi + \beta_i + \delta) \cdot \Psi(\theta - b_i) \quad (6.12)$$

Das zweidimensionale IRT-Modell lässt sich wiederum mit einem EM-Algorithmus schätzen.

6.5 Ländervergleich von vier Ländern in PIRLS 2011

Im folgenden Abschnitt soll ein Vergleich der Lesekompetenzleistungen in PIRLS 2011 anhand eines Testheftes für vier ausgewählte Länder Österreich (Ö), Deutschland (DEU), Frankreich (FRA) und Niederlande (NLD) vorgenommen und die Abhängigkeit der Ländermittelwerte in Abhängigkeit verschiedener Behandlungen der fehlenden Item Responses untersucht werden.

6.5.1 Daten

Für die Analyse sollen die Schülerantworten des Testheftes 13 in PIRLS 2011 („PIRLS Reader“) mit 35 Items (15 Multiple-Choice-Items mit 4 Antwortalternativen; 20 halboffene und offene Items) verwendet werden. Für das Testheft 13 liegen 968 österreichische, 809 deutsche, 901 französische und 802 niederländische Schüler/innen in der Stichprobe vor. Zur Vereinfachung der Analysen wurden alle polytomen Items dichotomisiert, wobei nur der höchste Punktwert bei einem Item zu einer Richtigantwort führt.

Deskriptive Analysen zeigen, dass die mittleren Missinganteile der Items zwischen den Ländern stark variierten (Ö: 11.2%, DEU: 7.9%, FRA: 13.6%, NLD: 2.7%). Diese Missinganteile waren bei den offenen Items deutlich stärker als bei den Multiple-Choice-Items ausgeprägt (z.B. Ö: offen – 17.7%, Multiple-Choice – 2.6%). Bei einer Bewertung der fehlenden Item Responses als Falschantwort erreichten österreichische Schüler/innen im Mittel 55.1% richtige Antworten und übertrafen damit leicht Frankreich (53.7%), erreichten aber signifikant niedrigere Ergebnisse als deutsche Schüler/innen (63.0%) und niederländische Schüler/innen (64.4%).

6.5.2 Analysen

Für die Ländervergleiche werden sechs verschiedene Item-Response-Modelle (im Folgenden als M1, ..., M6 bezeichnet) unter Berücksichtigung der Schülergewichte für das Testheft 13 berechnet. Für eine vereinfachte Beschreibung der Effekte der Länderreihenfolgen transformieren wir in jedem IRT-Modell erhaltene Fähigkeitsschätzungen so, dass österreichische Schüler/innen einen Mittelwert von 500 und eine Standardabweichung von 100 besitzen. Die Kompetenzwerte für Deutschland, Frankreich und die Niederlande wurden derselben Transformation unterworfen, so dass Ländervergleiche in allen Modellen immer relativ zur mittleren Leistung Österreichs vorgenommen wurden.

In Modell M1 – das dem offiziellen Vorgehen in PIRLS entspricht – werden in einem eindimensionalem Rasch-Modell mit vier Gruppen (den vier verschiedenen Ländern) fehlende Item Responses als falsch bepunktet. Im Modell M2 werden fehlende Item Response ignoriert, d.h. als Missing in der Likelihood-Schätzung betrachtet und damit ignoriert. In Erweiterung zu M2 wird in Modell M3 ein zweidimensionales Rasch-Modell mit den zwei latenten Variablen der Fähigkeit θ und der Response-Tendenz ξ spezifiziert. In Modell M4 werden im Pseudo-Likelihood-Ansatz des Rasch-Modells nur fehlende Item Response für die Multiple-Choice-Items mit vier Antwortalternativen mit dem Score $w_{pi} = 1/4$ bepunktet, während fehlende Item Responses bei offenen Items als falsch bepunktet werden (d.h. Score $w_{pi} = 0$). In den Modellen M1 bis M4 werden Itemschwierigkeiten und die Fähigkeitsverteilung in den vier Ländern simultan geschätzt.

Für die Modelle M5 und M6 werden in einer ersten Analyse durch eine Bewertung der fehlenden Item Response als Falschantwort gemeinsame Itemschwierigkeiten für alle vier Länder erhalten, die in einem zweiten Schritt fixiert werden. In Modell M5 wird der Pseudo-Likelihood-Ansatz (siehe 6.4.1) im Rahmen des Rasch-Modells eingesetzt, bei dem fehlende Item-Responses mit dem Score $w_{pi} = \rho P_{pi}(\theta_{p,M})$ versehen sind, wobei $\theta_{p,M}$ die Personenfähigkeit unter ausschließlicher Berücksichtigung der nichtfehlenden Item Responses bezeichnet. Die Personenfähigkeit $\theta_{p,M}$ wurde dabei aus der individuellen Posteriorverteilung simuliert. Die Ländermittelwerte werden in Abhängigkeit des Sensitivitätsparameters $\rho = 0, 0.01, \ldots, 0.99, 1$ berechnet. In Modell M6 wird die Erweiterung des zweidimensionalen Modells M3 durch die Modellierung des Ausfallprozesses die Abhängigkeit des Fehlens eines Item Responses vom Item selbst vorgenommen (siehe Abschnitt 6.4.2). Dieses Modell wird in Abhängigkeit des Sensitivitätsparameters $\delta = -10, -9.5, \ldots, -0.5, 0$ berechnet.

Für alle Item-Response-Modelle sollen die Ländermittelwerte Deutschlands, Frankreichs und der Niederlande im Vergleich zu Österreich verglichen werden.

Die gesamte Datenaufbereitung und Schätzung der Modelle fand in der Software R (R Core Team, 2014) statt. Für die Berechnung der Modelle M1 bis M3 wurde das R-Paket TAM (Kiefer et al., 2015) verwendet. Die in M4, M5 und M6 eingesetzten Modelle sind in der Funktion `rasch.mml2` des R-Paketes sirt (Robitzsch, 2015) implementiert.

6.5.3 Ergebnisse

In Tabelle 6.1 sind die Ländermittelwerte für alle Modelle M1 bis M6 dargestellt. Dabei fällt auf, dass zwischen den beiden Extrema der Missingbehandlung von Modell M1 (Miss-

ings werden als falsch bewertet) und Modell M2 (Missings sind ignorierbar) für Deutschland und Frankreich kleinere Unterschiede für die Differenz zum Österreich-Mittelwert entstehen (M1-DEU: 537.5, M2-DEU: 534.2; M1-FRA: 488.7, M2-FRA: 492.4), während die verschiedenen Modelle für die Niederlande zu deutlichen Differenzen führen (M1-NLD: 540.3, M2-NLD: 523.4). Dieser Befund ist aufgrund der deutlich geringeren Missinganteile bei den Item Responses (2.7%) der niederländischen Schüler/innen im Vergleich zum Missinganteil der Referenz der österreichischen Schüler/innen (11.2%) erklärbar.

Tabelle 6.1: *Mittelwerte für die Länder Österreich (Ö), Deutschland (DEU), Frankreich (FRA) und Niederlande (NLD). Der Mittelwert für Österreich wurde in allen Item-Response-Modellen auf 500 fixiert.*

Modell	Ö	DEU	FRA	NLD
M1: Missing = falsch	500	537.5	488.7	540.3
M2: Missing = ignorierbar	500	534.2	492.4	523.4
M3: 2-dim. Modell	500	534.9	492.5	524.8
M4: Pseudo-Likelihood (für Multiple-Choice-Items)	500	537.6	489.4	539.5
M5: Pseudo-Likelihood				
$\rho = 0$	500	537.3	488.9	539.9
$\rho = 0.3$	500	537.0	490.1	535.9
$\rho = 0.7$	500	535.9	491.8	529.5
$\rho = 1$	500	534.6	493.1	524.0
M6: 2-dim. Modellierung des Ausfallprozesses				
$\delta = -10$	500	538.0	489.1	540.7
$\delta = -1.5$	500	535.9	490.6	532.4
$\delta = -0.5$	500	535.1	491.5	528.0
$\delta = 0$	500	534.6	492.1	525.7

Die Modelle M5 und M6 variieren dabei die Sensitivitätsparameter ρ bzw. δ im Rahmen einer Sensitivitätsanalyse hinsichtlich verschiedener Annahmen an den Ausfallprozess der Item Responses. Praktisch bilden die Befunde aus Tabelle 6.1 die Spannweite der Ländermittelwerte zwischen den Extrema der Falschbewertung (Modell M1) und der Ignorierbarkeit (Modell M2). In Abbildung 6.1 erfolgt eine Darstellung der Ländermittelwerte in Abhängigkeit der Sensitivitätsparameter der Modelle M5 und M6. Die Ränge von Deutschland und der Niederlande tauschen dabei bei einem bestimmten Wert der Parameter ρ bzw. δ. Die Ländermittelwerte sind dabei stetige und monotone Funktionen der Sensitivitätsparameter.

6.6 Diskussion

Im Gegensatz zu Behauptungen der aktuellen Literatur (Pohl & Carstensen, 2013; Pohl et al., 2014) wurde anhand von Analysen in PIRLS 2011 gezeigt, dass das Ignorieren

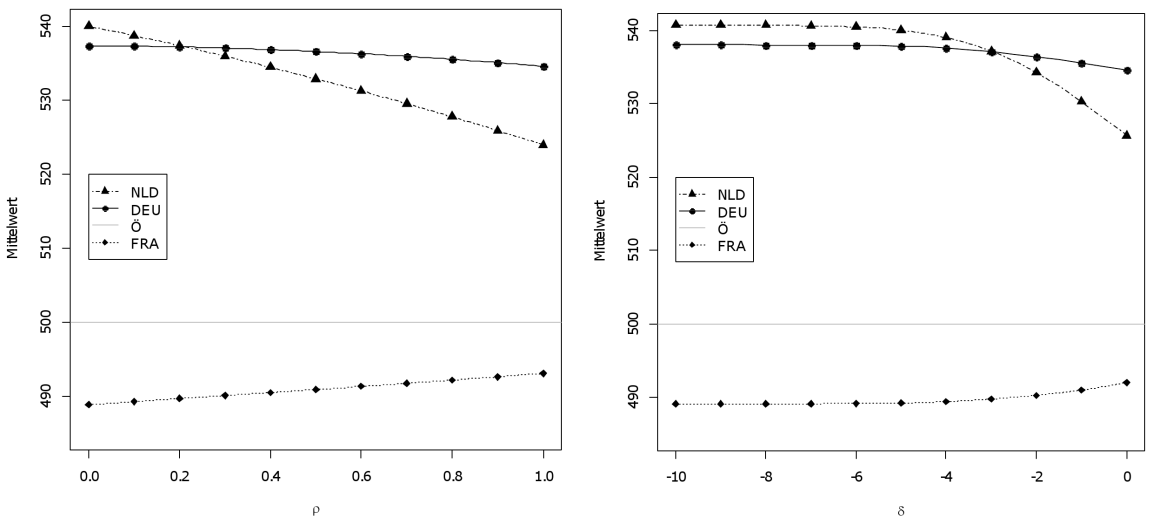

Abbildung 6.1: *Sensitivitätsanalysen für Ländermittelwerte der Länder Österreich (Ö), Deutschland (DEU), Frankreich (FRA) und Niederlande (NLD). Links: Pseudo-Likelihood-Schätzung (Modell M5) in Abhängigkeit des Sensitivitätsparameters ρ. Rechts: Zweidimensionales Modell M6 in Abhängigkeit des Sensitivitätsparameters δ*

fehlender Item Responses einen Einfluss auf zentrale Resultate einer Studie besitzt. Länderunterschiede bei verschiedenen Behandlungen fehlender Item Responses fallen dabei deutlich verschieden aus, was die Frage nach einer validen Analysemethode nach sich zieht.

Wir argumentierten, dass die in der Literatur auffindbare testtheoretisch basierte Kritik an einer Falschbewertung fehlender Item Responses (Rose, 2013; Pohl et al., 2014) bestenfalls auf einer intraindividuellen Interpretation von Wahrscheinlichkeiten beruht. In dieser Argumentation werden dann jedoch fälschlicherweise die Konzepte der aleatorischen Unsicherheit (das statistische Modell) und epistemischen Unsicherheit (das Zustandekommen der Daten betreffend) konfundiert. Es wurde ebenso aufgezeigt, dass Simulationsstudien für die Wahl der „richtigen Behandlung“ fehlender Item Responses keine Relevanz besitzen (siehe auch Rohwer, 2013). Da die Frage der Festlegung einer adäquaten Bepunktungsregel für die Missings nicht empirisch entscheidbar ist, wurden in diesem Beitrag zwei IRT-Modelle vorgeschlagen, die Sensititivitätsanalysen gegenüber der Annahme der Ignorierbarkeit fehlender Item Responses vornehmen. Das erste alternative eindimensionale IRT-Modell (Modell M5) basiert auf einem Pseudo-Likelihood-Ansatz, bei dem für fehlende Item Responses Punktwerte zwischen 0 und 1 zugelassen werden. Dieser Ansatz besitzt jedoch gegenüber dem zweiten alternativen zweidimensionalen IRT-Modell den Nachteil, dass die Bepunktung für die fehlenden Item Responses von einer ausschließlich auf den nichtfehlenden Items bestimmten weiteren Fähigkeit abhängen. Im zweiten Modell (Modell M6) wird neben der Fähigkeit auch eine dazu korrelierte Response-Tendenz angenommen. Dabei kann das Fehlen eines Item Responses auch vom unbekannten Item Response selbst abhängen. Rechnerisch ist das zweidimensionale Modell aufwändiger als das eindimensionale Modell und besitzt zusätzlich das Problem, dass keine Maximum-

Likelihood-Schätzung der Personenfähigkeit existiert (wenn δ verschieden von Null ist). Allerdings beruht in diesem Modell die Berechnung der Fähigkeit nicht wie bei Modell M5 auf einer zusätzlichen Ermittlung der Fähigkeit auf den nichtfehlenden Items.

Typischerweise werden die in diesem Artikel durchgeführten Sensitivitätsanalysen bei der Imputation fehlender Daten gegenüber komplexeren Modellspezifikationen wie Pattern Mixture Modellen (Pietsch, 2011) vorgezogen, wenn ein Bereich plausibler Analyseergebnisse unter einer Variation von plausiblen Annahmen an den Ausfallprozess vorgenommen wird (van Buuren, 2012).

Wird zusätzlich eine Priorverteilung über einen die Annahmen des Ausfallprozesses charakterisierenden Sensitivitätsparameter spezifiziert, so kann man im Rahmen einer Multi Model Inference Analyseergebnisse entsprechend theoretischer Vorannahmen gewichten (Siddique, Harel & Crespi, 2012), was häufig zu größeren Standardfehlern interessierender Modellparameter führt. Eine integrierte statistische Inferenz im Hinblick auf die Generalisierbarkeit hinsichtlich Personen, Items und verschiedener statistischer Modelle hat daher auch im Large-Scale Assessment Bedeutung (Robitzsch et al., 2011).

Im Rahmen unserer Analysen haben wir uns stets auf die Klasse der Rasch-Modelle beschränkt. Formal sind die Betrachtungen jedoch einfach auf komplexere IRT-Modelle wie das 2PL- oder das 3PL-Modell oder Modelle für polytome Daten übertragbar. Anstelle eines für alle Items gültigen Sensitivitätsparameters kann dieser separat für jedes einzelne Item oder für einzelne Itemgruppen (etwa Multiple-Choice Items und offene Items) spezifiziert werden. Explorative Analysen zeigen, dass man die Likelihood-Funktion auch in δ maximieren kann, d.h. der Typ des Datenausfallprozesses lässt sich empirisch bestimmen. Wenn nun allerdings der Ausfalltyp für Ländervergleiche länderspezifisch optimiert wird, so wird die Vergleichbarkeit stark in Frage gestellt. Dabei würde das psychometrische Modell in bedeutsamem Ausmaß die Skalierungsvorschrift bestimmen (siehe Brennan 2001 für eine kritische Auseinandersetzung).

Der vorgeschlagene Pseudo-Likelihood-Ansatz ermöglicht zusätzlich, Messfehler oder Unreliabilität bei der Erfassung der einzelnen Item Responses (etwa aufgrund eines Ratings bei einem offenem Item, das keine perfekte Reliabilität besitzt) zu berücksichtigen, in dem beobachtete Item Responses von 0 (falsch) bzw. 1 (richtig) auf nichtganzzahlige Scores modifiziert werden, die die Größe der Raterübereinstimmung abbilden soll. Alternativ zum Pseudo-Likelihood Ansatz kann allerdings auch mit dem statistischen Ansatz der Belief Functions nach Denoeux (2013) in IRT-Modellen gearbeitet werden, der auf anderen Modellannahmen der Likelihood-Funktion für die Schätzung beruht.

Neben der Betrachtung der Konsequenzen der Behandlung fehlender Item Responses für Personenfähigkeiten, hat die Analysestrategie durchaus auch Bedeutung für Itemparameterschätzungen (d.h. die Kalibrierung). Da offene Items meistens höhere Missinganteile als Multiple-Choice-Items besitzen, sind die Schwierigkeitsreihenfolgen der Items mit verschiedenen Itemformaten unter den Extrembehandlungen der Falschbewertung und der Ignorierbarkeit in vielen Large-Scale Assessments mit substanziellen Missinganteilen bedeutsam voneinander verschieden. Die Wahl der Methode der Itemparameterbestimmung hat allerdings Konsequenzen für die Verankerung von Items auf Kompetenzstufen und damit der Interpretation von Kompetenzenzwerten.

Die Konsequenzen einer Bewertung der fehlenden Item Responses als fehlende Itemantwort (und damit als ignorierbar) können vor allem bei Skalierungen in Längsschnittstudien

bedeutsam sein, wenn Missinganteile schulform- und klassenstufen- oder zeitpunktspezifisch (sowie in deren Interaktion) variieren. In IRT-Modellen mit ignorierbaren fehlenden Item Responses scheint in dieser Situation die Bedeutung einer ermittelten Fähigkeit unklar. Wird eine Fähigkeit in einer bestimmten Domäne in einem Test jedoch als Ausmaß der korrekten Beantwortung von Items einer vorgegebenen Itemmenge in einer vorgegebenen maximalen Testbearbeitungszeit definiert (wie dies beispielsweise in PIRLS/TIMSS oder PISA der Fall ist), so führt dies zu einer nichtzirkulären und aus unserer Sicht valideren Definition dieser Fähigkeit. Daraus folgt die Konsequenz, dass fehlende Item Responses nicht ignorierbar sein können.

Kapitel 7

Abschließendes Resümee

In diesem Kapitel werden wesentliche Aspekte der bisherigen Kapiteln zusammenfassen und systematisieren. Zunächst gehen wir in Abschnitt 7.1 auf die Modellierung von Positions-, Ermüdungs- und Kontexteffekten unter dem Gesichtspunkt von Modellabweichungen im Rasch-Modell ein. Im zweiten Abschnitt 7.2 wird die Bedeutung der Mehrebenenstruktur für Skalierungsmodelle untersucht. Im letzten Abschnitt 7.3 diskutieren wir Item-Response-Modelle und Faktormodelle unter der Perspektive einer unendlich groSSen Itempopulation (dem sog. Domain Sampling).

7.1 Modellierung von Positions-, Ermüdungs- und Kontexteffekten

In Kapitel 3 (Abschnitt 3.3) wurden exemplarisch IRT-Modelle für Positionseffekte, Ermüdungseffekte und Kontexteffekte (Bookleteffekte) diskutiert. Nachfolgend sollen diese Überlegungen etwas systematisiert werden und deren Konsequenzen für die Modellierung von Kompetenzen diskutiert werden. Da die diskutierten Effekte Verletzungen der Modellannahmen im Rasch-Modell darstellen, gehen wir im nächsten Abschnitt 7.1.1 allgemein auf Konsequenzen von Modellabweichungen im Rasch-Modell. In den weiteren Abschnitten werden die Folgerungen auf Positions-, Ermüdungs- und Kontexteffekte übertragen.

7.1.1 Modellabweichungen im Rasch-Modell

Zur Diskussion möglicher Verzerrungen bei der Anpassung fehlspezifizierter Modelle führen wir für das Rasch-Modell die Schreibweise latenter Item Responses ein. Mit der logistischen Linkfunktion Ψ gilt für das Rasch-Modell

$$\text{logit}\, P(X_{pi}=1|\theta_p) = \theta_p - b_i \qquad \text{bzw.} \qquad P(X_{pi}=1|\theta_p) = \Psi(\theta_p - b_i) \tag{7.1}$$

Den manifesten Item Responses X_{pi} möge nun latenter Item Response X^*_{pi} zugrunde liegen, wobei $X_{pi}=1$ genau dann, wenn $X^*_{pi}>0$ gilt (siehe De Boeck & Wilson, 2004, Kap. 1; Fox & Verhagen, 2010 für die Schätzung von Bayesianischen IRT-Modellen). Dabei lässt sich der latente Item Response schreiben als

$$X^*_{pi} = \theta_p - b_i + \epsilon_{pi} \tag{7.2}$$

wobei ϵ_{pi} ein logistisch verteiltes Residuum ist. In der Modellanpassung im Rasch-Modell wird die logistische Linkfunktion vorgegeben, was allerdings bedeutet, die Varianz der logistischen Verteilung (also $Var(\epsilon_{pi})$) zu fixieren (siehe De Boeck & Wilson, 2004). Die Äquivalenz der Beziehungen (7.1) und (7.2) ist dabei in der Literatur herausgearbeitet (De Boeck & Wilson, 2004, Kap. 1). Die Begründung basiert auf der Beziehung

$$P(X_{pi}=1) = P(X^*_{pi} > 0) = P(\theta_p - b_i > -\varepsilon_{pi}) \overset{(*)}{=} P(\varepsilon_{pi} < \theta_p - b_i) = \Psi(\theta_p - b_i) \quad (7.3)$$

Dabei ergibt sich $(*)$ aufgrund der Symmetrie der Verteilung von ε_{pi}. Damit ist gezeigt, dass die Parametrisierungen von X_{pi} und X^*_{pi} äquivalent sind. In den kommenden Überlegungen zur Behandlung von Positionseffekten und Kontexteffekten zeigt sich, dass die Schreibweise mit latenten Item Responses zu einfacheren Ableitungen führt.

Anstelle der Schreibweise (7.2) kann man mit $\sigma^2 = Var(\theta_p)$ eine Umparametrisierung vornehmen. Dafür definieren wir $\theta^*_p = \theta_p/\sigma$ (θ^*_p besitzt daher eine Varianz von 1) und $b^*_i = b_i/\sigma$. Die Standardabweichung σ wird dann als mittlere (oder gemeinsame) Itemladung interpretiert. In dieser Schreibweise ist

$$X^*_{pi} = \sigma(\theta^*_p - b^*_i) + \epsilon_{pi} \quad (7.4)$$

Wir merken an, dass sich die logistische Verteilung gut durch die Normalverteilung approximieren lässt, da für deren Verteilungsfunktionen Ψ und Φ näherungsweise die Beziehung $\Psi(x) \approx \Phi(Dx)$ mit $D = 1.702$ gilt (siehe De Boeck & Wilson, 2004, Kap. 1).

Wir nehmen nun an, dass Modellgleichung (7.4) nun noch durch einen Fehlerterm ν_{pi} erweitert wird, wobei wir $Var(\nu_{pi}) = \sigma^2_\nu$ annehmen:

$$X^*_{pi} = \sigma(\theta^*_p - b^*_i) + \nu_{pi} + \epsilon_{pi} \quad (7.5)$$

Diese Fehlerkomponente kann beispielsweise die Interaktion von Personen und Items in Testlets oder die Interaktion von Personen mit Items an bestimmten Positionen im Testheft indizieren. In vielen Anwendungen wird man nicht an einer expliziten Modellierung der Varianzquelle σ^2_ν interessiert sein. Daher fasst man bei der Anpassung eines Rasch-Modells beide Fehler in $\varepsilon_{pi} = \nu_{pi} + \epsilon_{pi}$ als unsystematische Fehlerquellen auf. Wegen $Var(\varepsilon_{pi}) = Var(\nu_{pi}) + Var(\epsilon_{pi}) = \sigma^2_\nu + D^2$ besitzt diese Behandlung allerdings Auswirkungen auf die extrahierte Traitvarianz σ. Dies möchten wir im Folgenden zeigen. Wir nehmen an, dass ε_{pi} wiederum logistisch verteilt ist, aber eine Varianz $\sigma^2_\nu + D^2$ besitzt. Dann können wir $\varepsilon_{pi} = \sqrt{\sigma^2_\nu/D^2 + 1} \cdot L$ mit einer logistisch verteilten Variablen L schrei-

ben. Wir erhalten (vgl. Ip, 2010; Tuerlinckx & De Boeck, 2004)

$$\begin{aligned} P(X_{pi}=1) &= P(X^*_{pi}>0) \\ &= \Psi(\sigma(\theta^*_p - b^*_i) + \nu_{pi}) \\ &= P\left(\sigma(\theta^*_p - b^*_i) + \nu_{pi} + \epsilon_{pi} > 0\right) \\ &= P\left(\sigma(\theta^*_p - b^*_i) + \varepsilon_{pi} > 0\right) \\ &= P\left(\sigma(\theta^*_p - b^*_i) + \sqrt{\sigma^2_\nu/D^2+1}\cdot L > 0\right) \\ &= P\left(\frac{\sigma}{\sqrt{\sigma^2_\nu/D^2+1}}\cdot(\theta^*_p - b^*_i) > -L\right) \\ &= \Psi\left(\sigma\cdot\frac{1}{\sqrt{\sigma^2_\nu/D^2+1}}\cdot(\theta^*_p - b^*_i)\right) \end{aligned} \tag{7.6}$$

Wenn man die Variable ν_{pi} „ausintegriert“ und daher zum Bestandteil der Fehlervarianz deklariert, so wird die extrahierte Traitvarianz im Rasch-Modell um den Faktor $1/\left(\sigma^2_\nu/D^2+1\right)$ verringert (siehe auch Snijders & Bosker, 2012, S. 307ff.). Dies erklärt den Befund, dass bei Gültigkeit des Testletmodells (bei dem der Term ν_{pi} Testletvarianz generiert) eine Anpassung des Rasch-Modells unter Ignorierung von Abhängigkeiten zu einer geringeren Traitvarianz führt, was gleichbedeutend mit einer verringerten mittleren Trennschärfe ist (DeMars, 2006; Tuerlinckx & De Boeck, 2001).

Die verringerte Traitvarianz scheint auf den ersten Blick konträr zu den Befunden in Abschnitt 5.2. Darin wurden die Konsequenzen positiver lokaler Abhängigkeiten in Faktormodellen mit identischer Linkfunktion (Normalverteilungsannahme) auf Basis von Kovarianzmatrizen diskutiert. Das Ignorieren der positiven Abhängigkeiten führte dabei zu einer überschätzten Traitvarianz und einer damit einhergehenden Überschätzung der Reliabilität. Im Faktormodell in Abschnitt 5.2 wäre das wahre Modell demzufolge $X^*_{pi} = \theta_p - b_i + \nu_{pi} + \epsilon_{pi}$, wobei die positiv lokalen Abhängigkeiten in der Variablen ν_{pi} abgebildet sind. Bei Ignorieren der lokalen positiven Abhängigkeit wird ein Teil der Varianz von ν_{pi} als wahre Varianz extrahiert. Dadurch verringert sich die Fehlervarianz und es resultiert die berichtete Unterschätzung dieser Varianzquelle. In Faktormodellen ist die Gesamtvarianz $Var(X^*_{pi})$ eines Items fixiert[1]. Im IRT-Modell mit logistischer Linkfunktion wird $Var(X^*_{pi})$ erst durch die Definition und Fixierung der Fehlervarianz auf $D^2 = 1.702^2 = 3.29$ definiert. Im IRT-Modell mit möglichen Testleteffekten (7.5) ist die Fehlervarianz durch $Var(\epsilon_{pi})$ bestimmt (und fixiert). Die Traitvarianz $\sigma^2 = Var(\theta_p)$ wird daher als relative Größe zu $Var(\epsilon_{pi})$ in Bezug gesetzt. Im IRT-Modell bei Ignorierung der Testleteffekte (7.6) wird die Fehlervarianz vergrößert, denn diese ist nun durch $Var(\nu_{pi})+Var(\epsilon_{pi})$ gegeben. Da im IRT-Modell die Varianz auf diese Größe fixiert wird, fällt die Traitvarianz bei Relativierung auf die (formal) nun größer ausfallende Fehlervarianz kleiner aus. Das Verhältnis der

[1]Bei Betrachtung tetrachorischer Korrelationen für dichotome Items wird $Var(X^*_{pi}) = 1$ gesetzt (vgl. Abschnitt 5.2).

Fehlervarianzen beider Modelle beträgt demzufolge

$$\frac{Var(\epsilon_{pi})}{Var(\nu_{pi}) + Var(\epsilon_{pi})} = \frac{D^2}{\sigma_\nu^2 + D^2} = \frac{1}{\sigma_\nu^2/D^2 + 1} \tag{7.7}$$

woraus der in (7.6) definierte Multiplikationsfaktor wiedergewonnen wird[2].

Der Forscher muss demnach in konkreten Anwendungen entscheiden, ob die Varianzquelle ν_{pi} Bestandteil der Fehlervarianz sein soll (und damit marginalisiert wird; *marginal modelling*) oder die Bestimmung der Fähigkeit explizit bedingt auf die ν_{pi} (*conditional modelling*) durchgeführt werden soll (Tuerlinckx & De Boeck, 2004). Formal lässt sich daher nicht über Bias sprechen, wenn bestimmte Modellabweichungen im Rasch-Modell, die sich in der Variablen ν_{pi} (ggf. systematisch) manifestieren, nicht explizit modelliert werden. Nur weil bestimmte Testheftpositionen oder Kontexte eine Varianz der Variablen ν_{pi} generieren, muss das nicht heißen, dass diese Varianzquellen im Interesse der Modellierung stehen. Im Gegenteil: Wenn eine Fähigkeit θ_p erst durch ein Bedingen auf eine Menge zusätzlicher Variablen ν_{pi} (z.B. mehreren Testleteffekten) definiert wird, so stellt sich die Frage der Bedeutung dieser Variablen. In den nächsten Abschnitten diskutieren wir konkrete Modellierungen für Positionen und Kontexte und wenden unsere Überlegungen darauf an.

7.1.2 Positionseffekte

In jüngerer Zeit werden vermehrt Ansätze zur Modellierung von Positionseffekten in nationalen und internationalen Schulleistungsstudien diskutiert (Debeer & Janssen, 2013, Debeer, Buchholz, Hartig & Janssen, 2014, Hartig & Buchholz, 2012, Weirich, Hecht & Böhme, 2014). Die Wahrscheinlichkeit einer richtigen Antwort von Person p für Item i an der Testposition k ist gegeben durch

$$\text{logit } P(X_{pik} = 1) = \theta_{pk} - b_{ik} \tag{7.9}$$

Hierbei bezeichnet θ_{pk} die Fähigkeit einer Person p an einer Testheftposition k und b_{ik} die Schwierigkeit von Item i an Position k. Dabei ist zu beachten, dass kein Item bei einer

[2] In einem Faktormodell mit stetigen Items oder der Schätzung des Faktormodells auf Basis tetrachorischer Korrelationen für dichotome Items diskutieren wir den Fall, dass die Items in T Testlets mit jeweils n_t Items vorliegen. Die Traitvarianz $Var(\theta)$ sei dabei σ^2, die als homogen über alle Testlets angenommene Testletvarianz sei ν^2. Verwendet als Schätzer der Traitvarianz $\hat{\sigma}^2$ die mittlere Kovarianz zwischen allen Items (was der unweighted least squares Schätzung entspricht), so erhält man

$$\hat{\sigma}^2 = \sigma^2 + \frac{1}{T} \cdot \frac{n_t - 1}{n_t - 1/T} \cdot \nu^2 \tag{7.8}$$

Daran erkennt man, dass für eine sehr große Anzahl von Testlets ($T \to \infty$) die Testleteffekte geringer ausgeprägt sind, d.h. $\hat{\sigma}^2$ ist eine asymptotisch erwartungstreue Schätzung für σ^2, wenn man die Anzahl der Items (und damit die Anzahl der Testlets) gegen Unendlich gehen lässt (siehe auch Stout, 1990 für ähnliche Überlegungen im Kontext essenzieller Eindimensionalität). Die alleinige Angabe der Größe der Testletvarianz (oder verschiedener Testletvarianzen) sagt daher im Allgemeinen wenig über die Konsequenzen lokaler stochastischer Abhängigkeiten für Itemparameterschätzungen und Varianzschätzungen aus (siehe auch Kapitel 5 dieser Arbeit).

Person zu verschiedenen Positionen vorgelegt wird. Dies unterscheidet diese Art der Modellierung von längsschnittlichen IRT-Modellen, wenn man die Positionen als Zeitpunkte interpretiert (z.B. te Marvelde, Glas, Van Landeghem & Van Damme, 2006).

Anstelle von (7.9) kann man – wie in Abschnitt 7.1.1 eingeführt – anstelle der dichotomen Item Responses X_{pik} latente Item Responses X^*_{pik} einführen. Das IRT-Modell ist dann durch

$$X^*_{pik} = \theta_{pk} - b_{ik} + \epsilon_{pik} \tag{7.10}$$

gegeben, wobei die Varianz von ϵ_{pik} auf die der logistischen Verteilung mit $Var(\epsilon_{pik}) = 3.29$ fixiert wird. Das IRT-Modell (7.10) ist ohne Nebenbedingungen im Allgemeinen nicht identifizierbar. Daher werden Untersuchungen zu Positions- und Ermüdungseffekten häufig als restringierte Modelle für die Fähigkeiten θ_{pk} und Itemschwierigkeiten b_{ik} verwendet, auf die wir nun näher eingehen.

Wird ein Rasch-Modell anstelle des Modells (7.10) angenommen, so ermittelt man *eine* Personenfähigkeit θ_p und eine Itemschwierigkeit b_i, so dass

$$X^*_{pik} = \theta_p - b_i + \underbrace{(\theta_{pk} - \theta_p)}_{=\nu_{pk}} + \underbrace{(b_{ik} - b_i)}_{=\nu_{ik}} + \epsilon_{pik} \tag{7.11}$$

erfüllt ist. Hierbei nehmen wir an, dass die Terme ν_{pk} bzw. ν_{ik} für jede Person p und jedes Item i den Erwartungswert (bzw. Mittelwert) Null besitzen. In der Anpassung des Rasch-Modells in (7.13) wird deutlich, dass bei Fixierung der logistischen Verteilung nun die gesamte Varianz

$$Var(\nu_{pk} + \nu_{ik} + \epsilon_{pik}) = Var(\nu_{pk}) + Var(\nu_{ik}) + Var(\epsilon_{pik}) \tag{7.12}$$

fixiert wird. In Analogie zu (7.6) verringert sich damit die extrahiere Traitvarianz um den Faktor $1/\left(\left[Var(\nu_{pk}) + Var(\nu_{ik})\right]/D^2 + 1\right)$. Wird demzufolge angenommen, dass Interaktionen von Personen und Positionen bzw. Items und Positionen unsystematische Varianz generieren (oder diese als unsystematische Varianzquelle aufgefasst werden), dann würde die Anpassung des Rasch-Modells selbst bei substanziellen Varianzquellen von ν_{pk} und ν_{ik} legitimiert. Die existierenden Varianzquellen können aber dazu führen, dass Standardfehler für Personenfähigkeiten θ_p und Itemschwierigkeiten b_i größer als unter der Annahme unabhängiger Fehler ausfallen (für Ergebnisse in der Generalisierbarkeitstheorie siehe Brennan, 2001a).

In der Literatur werden verschiedene Modellierungen vorgeschlagen. Wir starten zunächst mit Modellierungen auf der *Itemseite*

$$X^*_{pik} = \theta_p - b_{ik} + \epsilon_{pik} \tag{7.13}$$

Hierbei nehmen wir an, dass die Personenfähigkeit invariant über die Positionen t sein möge. Einige Ansätze modellieren die Positionseffekte (z.B. Alexandrowicz & Matschinger, 2008) itemunspezifisch, d.h.

$$b_{ik} = b + \gamma_k \tag{7.14}$$

mit festen Positionseffekten γ_k. Im Allgemeinen werden Items im Testheft mit wachsender Position schwieriger, weshalb γ_k typischerweise monoton in k ist. Es erscheint aber

unplausibel, dass alle Items in homogenem Ausmaß von Positionseffekten betroffen sind. Alternativ zu (7.14) könnte man auch

$$b_{ik} = b_i + \gamma_i(k - k_0) \tag{7.15}$$

mit einem itemspezifischen Positionseffekt γ_i postulieren, der linear ausgeprägt ist. Hierbei bezeichnet k_0 eine *Referenzposition*. Als Spezialfall von (7.13) und (7.15) kann man einen itemunspezifischen linearen Positionseffekt (siehe Debeer & Janssen, 2013) annehmen

$$b_{ik} = b_i + \gamma(k - k_0) \tag{7.16}$$

Die verschiedenen Modelle kann man empirisch an konkreten Datensätzen prüfen. Itemschwierigkeiten lassen sich dann von Positionseffekten trennen, wenn (mehrere) Items an mehr als einer Position im Test eingesetzt werden. In der konkreten Anwendung muss dabei entschieden werden, ob Positionseffekte Bestandteil der Modellierung sein sollen oder nicht. Liegt ein Item nur an einer Position im Testheft vor, so ist die Itemschwierigkeit jeweils nur auf die konkrete Position im Testheft zu interpretieren. Liegt ein Item an mehreren Positionen im Testheft vor, so ist die durch das Rasch-Modell bei Ignorierung möglicher Positionseffekte gewonnene Itemschwierigkeit b_i ein (stichprobengrößen-) gewichteter Effekt der positionsspezifischen Itemschwierigkeiten b_{ik}. Wenn ein Item in einem Test an zwei Positionen auftritt und dabei an Position 1 bei 40% und bei Position an 60% der Personenstichprobe administriert wird, so ist die Schwierigkeit b_i stärker von hinteren Positionen beeinflusst.

7.1.3 Ermüdungseffekte

Unter Ermüdungseffekten verstehen wir hier die Einführung weiterer Personenvariablen neben der Fähigkeit θ_p. Demzufolge konzentrieren wir uns im Folgenden auf die *Personenseite*. IRT-Modelle mit einem Ermüdungseffekt ξ_p (auch als Persistenz bezeichnet; Debeer & Janssen, 2013, Hartig & Buchholz, 2012)

$$\theta_{pk} = \theta_p + \xi_p(k - k_0) \tag{7.17}$$

Hierbei ist der individuelle Leistungsabfall linear modelliert. Aus Identifikationsgründen nimmt man dabei $E(\theta_p) = E(\xi_p) = 0$ an. In Modell (7.17) ist „Fähigkeit“ immer nur als eine Disposition zur Referenzposition k_0 zu sehen. Häufig wird die erste Position als Referenz gewählt. Aus empirischen Studien ist bekannt, dass die so definierte Fähigkeit negativ mit der Persistenz korreliert (Debeer & Janssen, 2013, Hartig & Buchholz, 2012), so dass eine höhere Fähigkeit mit einem geringeren individuellen Leistungsabfall einhergeht (vgl. Abschnitt 3.3 dieser Arbeit). Allerdings hängt diese Aussage explizit davon ab, dass die Fähigkeit durch die erste Position definiert ist. Wählt man in (7.17) eine andere Referenzposition k_0^*, so folgt für die umdefinierte Fähigkeit θ_p^* und die umdefinierte Persistenz ξ_p^*

$$\theta_{pk} = \theta_p^* + \xi_p^*(k - k_0^*) \tag{7.18}$$

Setzen wir (7.17) und (7.18) gleich, so folgt

$$\theta_p^* = \theta_p + \xi_p(k_0^* - k_0) \qquad \text{und} \qquad \xi_p^* = \xi_p \tag{7.19}$$

Damit ergibt sich für die Kovarianz

$$Cov(\theta_p^*, \xi_p^*) = Cov(\theta_p + \xi_p(k_0^* - k_0), \xi_p) = Cov(\theta_p, \xi_p) + (k_0^* - k_0)Var(\xi_p) \tag{7.20}$$

Wenn die Kovarianz $Cov(\theta_p, \xi_p)$ mit $k_0 = 1$ negativ ausfällt, so wird für eine mittlere Testposition $k_0^* > k_0$ die Kovarianz $Cov(\theta_p^*, \xi_p^*)$ (und damit auch die Korrelation) höher (und weniger negativ bzw. ggf. auch positiv) ausfallen. Eine berichtete Korrelation zwischen Fähigkeit und Persistenz sollte daher immer auch die Information enthalten, ob die Fähigkeit durch die erste Testposition oder als Fähigkeit über alle Testpositionen hinweg definiert ist. Ich tendiere dazu, als „Referenzpunkt" für die Fähigkeit die mittlere Testposition zu wählen, um die mittlere Leistung im gesamten Test abzubilden.

Beim Einsatz des Modells (7.17) wählt man zusätzlich auch häufig itemunspezifische lineare Positionseffekte wie in (7.16). Damit wird folgendes lineares „Wachstumskurvenmodell" spezifiziert:

$$X_{pik}^* = \theta_p + \xi_p(k - k_0) - b_i - \gamma(k - k_0) + \epsilon_{pik} = \theta_p - b_i + (\xi_p - \gamma)(k - k_0) + \epsilon_{pik} \tag{7.21}$$

Das Rasch-Modell wird damit um einen linearen personenspezifischen Trend (der Ermüdung) und einen linearen itemspezifischen Positionseffekt ergänzt. Flexiblere Trendschätzungen wären durch Anwendung eines *curves-of-factor* Modells (vgl. Weirich, Penk, Hecht, Roppelt & Böhme, submitted) möglich, in der der lineare Trend durch einen nichtlinearen Trend $\gamma(k) = \gamma_k$ ersetzt wird. Dabei ist

$$X_{pik}^* = \theta_p - b_i + (\xi_p - 1) \cdot \gamma_k + \epsilon_{pik} \tag{7.22}$$

wobei $\gamma_{k_0} = 0$ gesetzt wird. Wie in Modell (7.21) hängt auch in diesem Modell die Fähigkeit wiederum von der Definition der Referenzposition ab. IRT-Modelle für Ermüdungseffekte unter Berücksichtigung der Nestung von Schüler in Schulen diskutieren Albano (2013) und Debeer et al. (2014).

Formal kann man jede parametrische Reduktion der personenspezifischen Effekte θ_{pk} einsetzen, die auch der Analyse von Längsschnittdaten üblich ist (Zimmerman & Nunez-Anton, 2009). Ein Modell mit positionsspezifischen Fähigkeiten ist in Weirich et al. (submitted) diskutiert. Selbst wenn diese IRT-Modelle eine bessere Modellpassung aufweisen, so bleibt ungeklärt, wie eine „unverzerrte" Schätzung *einer* Fähigkeit ableitbar ist. Von Interesse wird eine gewichtete Fähigkeit $\tilde{\theta}_p = \sum_k w_k \theta_{pk}$ mit Gewichten w_k sein. Wenn die positionsspezifischen Fähigkeiten ungefähr dieselbe Varianz aufweisen, könnte eine Gleichgewichtung der Fähigkeiten θ_{pk} plausibel sein. Wenn die Generalisierung auf die gesamte Testzeit von Interesse ist, so könnte eine Schätzung mit Hilfe des Rasch-Modells sinnvoll sein, da damit eine Gleichgewichtung der Items zu allen Zeitpunkten gegeben ist. Liegt ein Außenkriterium V vor, so könnte man die Gewichte w_k mit einer kanonischen Korrelationsanalyse so bestimmen, dass die ermittelte Fähigkeit $\tilde{\theta}_k$ mit V eine maximale Korrelation aufweist (vgl. Ballou, 2009; Cunha et al., 2010).

7.1.4 Bedeutung für Gruppenvergleiche

Im Folgenden sollen die Überlegungen zu Positions- und Ermüdungseffekten im Hinblick auf Gruppenvergleiche diskutiert werden. Zur Illustration mögen dabei für zwei Personengruppen g und h mittlere Fähigkeitswerte θ_g und θ_h sowie mittlere Persistenzen ξ_g und ξ_h

für die Referenzposition k_0 vorliegen. Beispielsweise könnte g die österreichischen Schüler und h die slowenischen Schüler bezeichnen. Unter der Referenz der ersten Testheftposition $k_0 = 1$ mögen österreichische Schüler besser als slowenische Schüler abschneiden, also $\theta_g > \theta_h$. Die mittlere Persistenz solle nun aber bei slowenischen Schülern höher sein, d.h. $\xi_g < \xi_h$. Würde man nun alternativ die mittlere Fähigkeit für die Position k_0^* als Referenz berichten, so ergibt sich $\theta_g^* = \theta_g + \xi_g(k_0^* - k_0)$ sowie $\theta_h^* = \theta_h + \xi_h(k_0^* - k_0)$. Der Fähigkeitsunterschied zwischen beiden Gruppen beträgt dann $\theta_g^* - \theta_h^* = \theta_g - \theta_h + (\xi_g - \xi_h)(k_0^* - k_0)$, d.h. diese Differenz hängt von der Wahl der Referenzposition ab. Das illustrierte Beispiel besitzt auch in internationalen Vergleichsstudien wie PISA Relevanz (Le, 2009; Hartig & Buchholz, 2012).

Ich argumentiere, dass Gruppenunterschiede immer vor dem Hintergrund des Testkonzepts (und damit auch unter dem gewählten Testdesign) zu interpretieren sind. Wenn also der Leistungsunterschied von österreichischen und slowenischen Schülern in den ersten 60 Minuten der PISA-Testung anders als in den zweiten 60 Minuten ausfällt, dann schränkt dies nicht die Gültigkeit der dokumentierten Leistungsunterschiede auf Basis der Gesamtzeit von 120 Minuten ein. Wie in Abschnitt 7.1.3 beschrieben, könnte man also eine empirisch oder normativ getriebene Gewichtung der positionsspezifischen Fähigkeiten vornehmen. Ob diese von der Gleichgewichtung im Allgemeinen verschiedene Neugewichtung dann noch die „Leistung in PISA“ abbildet, scheint allerdings fraglich.

7.1.5 Kontexteffekte und Bookleteffekte

Im Folgenden möchten wir Kontexteffekte genauer von Positionseffekten abgrenzen. Bisher assoziieren wir mit Positionseffekten die veränderliche Itemschwierigkeit b_{ik} von Item i von der Position k. Items können jedoch in verschiedenen Reihenfolgen in Testheften (*Kontexten*) auftreten. Wir bezeichnen im Folgenden alle Effekte, die auf das *gemeinsame Auftreten* von Items in einem Kontext zurückführbar sind, als *Kontexteffekte*. Als typisches Beispiel eines Kontextes betrachten wir die Vorlage von Items in verschiedenen Testheften (Booklets; vgl. Hecht, Weirich, Siegle & Frey, 2015).

Zur Formalisierung betrachten wir nun (wiederum stetige Versionen der) Itemantworten X_{pikc}^* von Person p auf Item i zur Position k in Kontext c:

$$X_{pikc}^* = \theta_{pkc} - b_{ikc} + \epsilon_{pikc} \tag{7.23}$$

Dabei wird für jede Person p nur genau ein Kontext (z.B. ein Testheft) administriert. Die Fähigkeiten θ_{pkc} und Itemschwierigkeiten b_{ikc} können zunächst zwischen Kontexten und Positionen variieren. Typischerweise wird man wiederum aus Identifikationsgründen für die Fähigkeiten θ_{pkc} Erwartungswerte von Null fordern. Dies entspricht der typischen Situation eines Booklet-Designs, in dem man unter der Annahme äquivalenter Gruppen Testhefte Personen zufällig zuteilt. In empirischen Studien kann dabei durchaus die Itemschwierigkeit auch bei gleicher Position im Testheft zwischen verschiedenen Testheften variieren, was durch die Verwendung des separaten Parameters b_{ikc} indiziert wird.

In Tabelle 7.1 ist ein querschnittliches Testdesign abgebildet, das in ähnlicher Weise häufig im Large-Scale Assessment eingesetzt wird (Rutkowski, Gonzales, von Davier & Zhou, 2014). Items werden dabei in Itemblöcken (A, B und C) gruppiert, wobei jedes Item in genau einem Itemblock auftritt. Testhefte bestehen dabei aus einer Sequenz von

Itemblöcken, die einer systematischen Anordnung folgen (siehe auch Frey, Hartig & Rupp, 2009). Im Design in Tabelle 7.1 tritt jeder der Itemblöcke genau dreimal an der ersten und dreimal an der zweiten Position auf. Beispielsweise kann ein Item i im Itemblock A auftreten. Die Itemschwierigkeit b_i des Items i wird dabei von der Testheftposition abhängen. Erwartungsgemäß sollte die Schwierigkeit von i in Testheft 2 (TH2) größer als in Testheft 1 (TH1) sein, da in TH2 das Item an späterer Position im Testheft vorgelegt wird. Neben der Position könnte allerdings auch der Kontext relevant sein. In den Testheften 2, 4 und 8 ist das Item i im Itemblock A jeweils an zweiter Position administriert. Die Itemschwierigkeit kann aber in Abhängigkeit des vorhergehenden Blockes variieren. In Testheft 8 könnte dabei an der ersten Position ein Itemblock aus einem anderen Test oder ein Schülerfragebogen vorgelegt werden. Dann könnte es sein, dass Item i leichter wird, wenn kein Itemblock desselben Tests an der ersten Position administriert wird.

Tabelle 7.1: *Querschnittliches Design mit 12 Testheften und zwei Testpositionen*

Testheft	Pos1	Pos2
TH1	A	B
TH2	B	A
TH3	A	C
TH4	C	A
TH5	B	C
TH6	C	B
TH7	A	–
TH8	–	A
TH9	B	–
TH10	–	B
TH11	C	–
TH12	–	C

Anmerkung: Die Einträge „–“ kennzeichnen eingesetzte Itemblöcke aus einem anderen Test. Alternativ könnte der eingesetzte Test in Testheft 7 nur verkürzt worden sein oder in Testheft 8 zu einer späteren Uhrzeit eingesetzt werden.

Unter der Annahme äquivalenter Gruppen wären alle Itemschwierigkeiten b_{ikc} schätzbar. Für praktische Modellierungen werden jedoch im Allgemeinen sparsamere Parametrisierungen eingesetzt. PISA führt in der Skalierung der Fähigkeitswerte Bookletparameter δ_c ein (OECD, 2014), so dass das IRT-Modell die Form

$$X^*_{pikc} = \theta_p - b_i - \delta_c + \epsilon_{pikc} \tag{7.24}$$

besitzt. Formal wird also die Gleichheit $b_{ikc} = b_i + \delta_c$ unterstellt. Das IRT-Modell (7.24) „ignoriert“ daher Positionseffekte und wird im Allgemeinen fehlspezifiziert sein. Das Modell ist allerdings relativ sparsam und adjustiert die Schätzungen der Fähigkeitswerte im Ausmaß der mittleren Schwierigkeit des Booklets c. Interaktionen von Items und Positionen sowie Items und Kontexten werden demzufolge in der Fehlervarianz von ϵ_{pikc} abgebildet.

Der Bookletparameter δ_c kann jedoch auch als ein mittlerer Effekt interpretiert werden. Es sei angenommen, dass bei einem balancierten Booklet-Design (alle Items an allen Positionen in „ausgewogenen“ Kontexten) die Itemparameter b_i geschätzt werden können. Für ein konkretes Booklet c sind die Itemparameter b_{ikc} bestimmbar, so dass insgesamt I_c Parameter im Booklet mit I_c Items vorliegen. Damit lassen sich Abweichungen $\nu_{ikc} = \nu_{ic} = b_{ikc} - b_i$ bestimmen. Im IRT-Modell (7.24) mit Bookleteffekten nimmt man dann $\nu_{ic} = \delta_c + e_{ic}$ an, d.h. die Abweichungen werden *im Mittel* durch den Bookletparameter δ_c approximiert und die Approximationsfehler e_{ic} mitteln sich aus. Formalisiert bedeutet dies

$$X^*_{pikc} = \theta_p - b_{ikc} + \epsilon_{pikc} = \theta_p - b_i - \delta_c + \underbrace{(-e_{ic}) + \varepsilon_{pikc}}_{=\tilde{\varepsilon}_{pikc}} \tag{7.25}$$

In dieser Schreibweise wird deutlich, dass residuale Effekte der Interaktion von Item und Kontext nunmehr Bestandteil der Fehlervarianz sind. Im Mittel muss dies zu keinem „Bias“ führen, wenn die Größe der Varianzen von e_{ic} nicht stark zwischen den Testheften schwankt. Die Varianz der Personenfähigkeit θ_p wird allerdings bei dieser Art der Nichtmodellierung (ggf. leicht) geringer geschätzt als bei einer konkreten Modellierung.

Anstelle der Verwendung des Modells (7.24) mit Bookleteffekten könnte argumentiert werden, dass ein Modell unter Berücksichtigung von Itempositionseffekten eingesetzt werden könnte, d.h. es wird

$$X^*_{pikc} = \theta_p - b_{ikc} + \epsilon_{pikc} = \theta_p - b_{ik} + \varepsilon_{pikc} \tag{7.26}$$

mit einer geeigneten Modellierung von b_{ik} betrachtet. Dabei kann beispielsweise eines der Modelle (7.14), (7.15) oder (7.16) zum Einsatz kommen. In (7.26) wird demzufolge die Approximation $b_{ikc} = b_{ik} + e_{ikc}$ mit Residuen e_{ikc} verwendet. Hierbei ist anzumerken, dass die Approximation b_{ik} nun *nicht* mehr vom Kontext, sondern nur von der Position abhängt. Für die Schätzung dieses Parameters wird man typischerweise auch andere Booklets heranziehen (wie z.B. mit der Parametrisierung $b_{ik} = b_i + \gamma_k$). Ich argumentiere, dass mit dieser Modellierung nicht gesichert ist, dass in Anwendungen der mittlere Approximationsfehler e_{ikc} gleich Null ist. D.h. die Zusammensetzung des Booklets (Itemreihenfolge, Verteilung der Itemschwierigkeiten im Booklet) kann neben den Positionseffekten nichtvernachlässigbare Konsequenzen besitzen. In diesem Sinne ist die Modellierung mit Bookleteffekten (7.25) robuster als die Modellierung (7.26) ausschließlich mit Positionseffekten.

Bedeutung von Kontexteffekten im Large Scale Assessment

Wir konkretisieren nun unsere Überlegungen am Testdesign in Tabelle 7.1. Durch die Anlage dieses Designs wird das Vorgehen in internationalen Studien (wie PISA oder PIRLS/TIMSS) oder nationalen Studien (deutscher Ländervergleich) illustriert. Typischerweise werden in diesen Studien Testhefte den Schülern zufällig zugeordnet. Demzufolge wäre dann die Annahme äquivalenter Gruppen plausibel und alle Unterschiede von Testleistungen zwischen den Testheften sind auf Unterschiede in Testheftschwierigkeiten und nicht auf Unterschiede in Schülerfähigkeiten zurückzuführen. Die wenigsten Annahmen werden demzufolge mit einem *Equating* (Kolen & Brennan, 2004) getroffen, in dem angenommen wird, dass die Fähigkeitsverteilungen zwischen allen Heften identisch

ist. Dies ist (im Normalverteilungsfall) dazu äquivalent, die Fähigkeitsverteilung auf eine Standardnormalverteilung zu fixieren und die Testhefte separat zu skalieren.

Wenn (wie in PISA) das IRT-Modell mit Bookleteffekten (7.25) eingesetzt wird, dann nimmt man an, dass die verschiedenen Testhefte (mit verschiedenen mittleren Positionen der Items und verschiedenen Kontexten) Fähigkeitsverteilungen mit verschiedenen Mittelwerten bei einer Skalierung mit dem Rasch-Modell und Itemschwierigkeiten b_i generieren würden. Da aber von äquivalenten Gruppen ausgegangen werden soll, werden mittlere Fähigkeitsunterschiede adjustiert. Die Skalierung mit dem Rasch-Modell und Bookleteffekten (7.25) erfolgt dann aber mit der Annahme, dass die extrahierte Traitvarianz für alle Testhefte dieselbe ist. Dies würde aber für das Testdesign in 7.1 heißen, dass die Traitvarianz für die ersten sechs Testhefte mit Items an zwei Positionen genauso groß wäre wie für die Testhefte 7, 9 und 11, in denen Items nur an der ersten Position vorliegen. Diese Annahme scheint vor der Hintergrund der Existenz von Ermüdungseffekten (siehe Modell (7.17)) nicht plausibel und kann empirisch getestet werden.

In einem Testdesign wie in Tabelle 7.1 wird man aber davon ausgehen können, dass die extrahierten Traitvarianzen testheftspezifisch ausfallen. Man könnte dann einwenden, dass mit allen Testheften nicht „dieselbe Fähigkeit“ gemessen wird. Die mit einem Testdesign aus allen Testheften ermittelte Traitvarianz könnte aber auch designbasiert interpretiert werden. Dabei ist die gesamte Traitvarianz die gewichtete Summe der Varianzen der einzelnen Testhefte, wobei die Gewichtung der Testhefte durch die Anlage des Designs oder eine nachträgliche Stichprobengewichtung vorgenommen wird. Im Kontext von Kapitel 4 können die verschiedenen Testhefte (Kontexte) als „Modelle“ interpretiert werden, die entsprechend eines Sampling-Plans gezogen (bzw. gewichtet) werden. „Unverzerrte“ Parameterschätzer entstehen dann dadurch, dass die verschiedenen möglichen Kontexte in einem Testdesign repräsentativ abgebildet werden. Formal lassen sich diese Überlegungen auf einen beliebigen statistische Parameter γ (wie eine Varianz, ein Quantil oder ein Regressionskoeffizient) übertragen. Mit jedem Testheft c ist ein Parameter γ_c assoziiert. Schätzt man den Parameter γ auf Basis aller Testhefte, so ergibt sich eine gewichtete Schätzung mit Testheftgewichten w_c (die häufig den Stichprobenumfängen der Testhefte

im Verhältnis zu allen Testheften entsprechen)[3].

Bezug der Fähigkeit auf die erste Testheftposition

Für die zwei Positionen im Testdesign in Tabelle 7.1 werden daher im Allgemeinen zwei korrelierte Fähigkeiten θ_{p1} und θ_{p2} existieren. Man kann diese Fähigkeiten an den zwei Positionen als Testlets (siehe Kapitel 5) interpretieren. Skaliert man mit dem Rasch-Modell alle Items an beiden Positionen, so „ignoriert“ man diese „Abhängigkeiten“. Alternativ könnte man einen testheftpositionsübergreifenden Faktor θ'_p höherer Ordnung etablieren, in dem $\theta_{pk} = \theta'_p + E_{pk}$ angenommen wird (Rijmen, 2010). Die diskutierten IRT-Modelle mit Ermüdungseffekten (7.17) definieren Fähigkeit (praktisch) als Leistung an der ersten Testposition, d.h. es wird $\theta_{p1} = \theta'_p$ und $\theta_{p2} = \theta'_p + \xi_p$ definiert. Bei genau zwei Testpositionen ist dieser Ansatz aber dazu äquivalent zwei Fähigkeiten θ_{p1} und θ_{p2} anzunehmen. Im Testdesign in Tabelle 7.1 fällt auf, dass dann für die Testhefte 8, 10 und 12 nur Items an der zweiten Position vorliegen und somit nur θ_{p2} und nicht θ_{p1} bei alleiniger Betrachtung dieser Testhefte identifiziert ist. Für Schüler mit diesen Testheften liegen demzufolge keine Items vor, die die Fähigkeit zum ersten Zeitpunkt messen. Die Extrapolation auf den ersten Zeitpunkt kann bei Kenntnis des Zusammenhanges von θ_{p1} und θ_{p2} aus den Testheften TH1 bis TH6 geschlossen werden. Die in PISA für mehrere Kompetenzbereiche eingesetzte Plausible Value Technik (OECD, 2014; Mislevy, 1991; von Davier & Sinharay, 2014) kann genutzt werden, um die Inferenz von θ_{p2} auf θ_{p1} durchzuführen. Damit unterscheidet sich dieses diskutierte Verfahren nicht vom Vorgehen in PISA, Plausible Values der Lesekompetenz in Testheften mit Hilfe mathematischer Kompetenz und naturwissenschaftlicher Kompetenz zu berechnen, wenn in einem Testheft gar keine Items zur Lesekompetenz administriert wurden.

Insgesamt argumentieren wir, dass (mindestens in großen Stichproben) die Verwendung von Modellen mit Bookleteffekten zur Gewinnung von Itemschwierigkeiten und Personenfähigkeiten zu präferieren ist. Auch wenn eine Modellierung der Fähigkeit an

[3]Eine Maximum Likelihood Schätzung für γ maximiert dann die Summe der Log-Likelihood-Terme aller Testhefte c, formal ist

$$l(\gamma) = \sum_c w_c \cdot \log L(\boldsymbol{X}_c|\gamma) = \sum_c w_c \cdot l_c(\gamma) \tag{7.27}$$

In jedem Testheft c wäre dabei γ_c der optimale geschätzte Parameter, d.h. es gilt mit $l' = \frac{\partial l_c}{\partial \gamma}$ die Beziehung $l'_c(\gamma_c) = 0$. Für die Daten des gesamten Tests gilt die Bestimmungsgleichung $l'(\gamma) = \sum_c w_c \cdot l'_c(\gamma) = 0$. In diesem Sinne ist also die „mittlere Schätzung“ γ über alle Testhefte zu verstehen.

Schreibt man l_c mit Hilfe der Taylorformel als $l_c(\gamma) \approx h_c - a_c(\gamma - \gamma_c)^2$ (wobei $a_c > 0$ die Präzision der Schätzung γ_c charakterisiert) und setzt in (7.27) ein, so ergibt sich

$$l'(\gamma) = -2 \sum_c w_c a_c (\gamma - \gamma_c) = 0 \tag{7.28}$$

Führt man relative Gewichte $\tilde{w}_c = w_c a_c / \left(\sum_c w_c a_c\right)$ für die Testhefte c ein, so ergibt sich

$$\gamma = \sum_c \tilde{w}_c \gamma_c \tag{7.29}$$

Der Schätzer γ ist damit eine mit $\tilde{w}_c = w_c a_c$ gewichtete Summe der testheftspezifischen Schätzer γ_c.

der ersten Testheftposition von Interesse ist, so spricht dies nicht zwangsläufig für den Einsatz von Modellen mit interindividueller Fähigkeit und interindividueller Ermüdung (Debeer & Janssen, 2013), da diese Modelle mögliche Kontexteffekte ignorieren. Im Rahmen eines modellbasierten Vorgehens kann die Ermittlung der Verteilung der Fähigkeit zur ersten Position auch mit einem Modell mit Bookleteffekten erfolgen. Wenn die Mittelwerte der Fähigkeiten zu beiden Testpositionen und in allen Testheften aufgrund der Annahme äquivalenter Gruppen gleich sein sollen, so bietet sich folgendes allgemeineres IRT-Modell mit bookletspezifischen Positionseffekten δ_{kc} an

$$X^*_{pikc} = \theta_{pk} - b_i - \delta_{kc} + \epsilon_{pikc} \tag{7.30}$$

Wenn in manchen Testheften keine Items zu einer bestimmten Position vorliegen, so wird man den Parameter δ_{kc} gleich Null setzen. Der mehrdimensionale Fähigkeitsvektor $(\theta_{pk})_k$ kann mit Hilfe der Plausible Value Technik gewonnen werden. Anstelle der Definition der Fähigkeit durch die erste Testposition präferiere ich die Verwendung einer durch alle Testpositionen gewichteten Fähigkeit (siehe Abschnitt 7.1.3).

Konfundierung von Ermüdungs- und Positionseffekten mit Testleteffekten

Gerade bei der Erfassung von Lesekompetenz mit Testlets (wie in PISA, siehe Monseur, Baye, Lafontaine & Quittre, 2011) wird die Bestimmung von Ermüdungseffekten mit Testleteffekten konfundiert sein. Berichtete Testleteffekte könnten demzufolge nicht der Abhängigkeit von Items zu einem gemeinsamen Stimulus, sondern eher einer Interaktion von Person und Testposition zuzuschreiben. Besteht ein Test aus einer (festen) Abfolge einzelner Testteile (etwa Textsorten wie im TestDaF, siehe Eckes, 2015a), so würde eine Kontrolle der Testletvarianz möglicherweise der Kontrolle der Varianz individueller Ermüdungseffekte sein.

Umgekehrt könnten aber auch berichtete Ermüdungseffekte wie in PISA teilweise Testleteffekte sein (etwa in Debeer & Janssen, 2013). Entscheidend ist, welche Reliabilität man dabei einer aus einem IRT-Modell ermittelten Fähigkeit zuschreibt. Dies führt dazu, dass wir bei der Beurteilung der Geeignetheit von IRT-Modellen für Leistungstests tendenziell davon ausgehen, dass lokale stochastische Unabhängigkeit *keine* zu testende Modellannahme darstellt, sondern vielmehr als eine Setzung unter der Annahme einer hypothetischen Testwiederholung (Brennan, 2001c) entsteht. Wir greifen diese Überlegungen in Abschnitt 7.3 im Kontext des Domain Samplings wieder auf.

7.1.6 Bedeutung von Positions-, Ermüdungs- und Kontexteffekten in Längsschnittanalysen

Positionseffekte und Kontexteffekte erlangen auch in längsschnittlichen Analysen Bedeutung, wenn nicht dieselben Testinstrumente zu den Zeitpunkten eingesetzt werden. Ein typisches Testdesign ist in Tabelle 7.2 dargestellt. Zum ersten Zeitpunkt (T1) wird ein Testheft mit den Itemblöcken A und B administriert, wobei B die schwierigeren Items entält. Zum zweiten Zeitpunkt (T2) werden die Itemblöcke B und C eingesetzt, wobei B nun gegenüber C die leichteren Items beinhaltet. Diese Testabfolge wird motivations-

psychologisch begründet, in dem die leichteren Items zu Beginn des Tests administriert werden sollten.

Tabelle 7.2: *Längsschnittdesign für zwei Zeitpunkte (T1, T2) und jeweils einem Testheft (TH1, TH2) mit jeweils zwei Itemblöcken*

Zeitpunkt	Pos1	Pos2
T1 (TH1)	A	B
T2 (TH2)	B	C

Offensichtlich zeigt das Testdesign in Tabelle 7.2, dass die Verbindung zwischen den beiden Zeitpunkten nur mit Hilfe der Ankeritems im Itemblock B erfolgen kann. Die Ankeritems werden allerdings nicht an derselben Testposition administriert. Es könnte nun argumentiert werden, dass Positionseffekte eine Verzerrung in der Erfassung der mittleren Veränderung verursachen. Die mittleren Lösungshäufigkeiten von Items in Block B werden dabei typischerweise niedriger als an der ersten Position ausfallen. Da der Itemblock B aber zu T2 an Position 1 administriert wurde, wäre die beobachtete Leistungsdifferenz zwischen den beiden Zeitpunkten auf Basis der Items in B allerdings höher als wenn man die (hypothetische) Itemschwierigkeit für Items in B für T1 zur ersten Position berechnen würde.

Positionseffekte ließen sich allerdings mit dem Design in Tabelle 7.3 prüfen. Zu jedem Zeitpunkt werden nun zwei Testhefte vorgelegt. Dabei werden jeweils die Positionen der Itemblöcke vertauscht.

Tabelle 7.3: *Längsschnittdesign für zwei Zeitpunkte (T1, T2) und jeweils zwei Testheften*

Zeitpunkt	Pos1	Pos2
T1 (TH1)	A	B
T1 (TH2)	B	A
T2 (TH3)	B	C
T2 (TH4)	C	B

Gemäß unserer Überlegungen würde man nun die direkte Veränderung in den Items in Block B mit den Testheften TH2 und TH3 messen können. Die ermittelte Veränderung wäre dann eine Aussage über die Veränderung an der ersten Testposition. Man könnte die mittlere Veränderung ebenso über die Items in B an der zweiten Testposition mit den Testheften TH1 und TH4 ermitteln. Dann wären allerdings der Kontext nicht konstant gehalten, da verschiedene Itemblöcke an der ersten Testposition eingesetzt wurden.

Man kann sich fragen, ob Testdesigns wie in Tabelle 7.2, in denen die Ankeritems nicht an denselben Positionen vorgegeben werden, für die Bestimmung einer „wahren“ Veränderung von T1 nach T2 geeignet sind. Dafür sollte man zwei Konstellationen unterscheiden. Wenn die Itemblöcke A, B und C in Tabelle 7.2 als austauschbar betrachtet werden und sich diese Blöcke strukturell im Hinblick auf das Konstrukt nicht unterscheiden, so kann man von einer verzerrten Schätzung ausgehen, da Schüler zu T1 bei Vorlage des Itemblocks B an der zweiten Position gegenüber der Vorlage von B an der ersten Position zu T2 einen Nachteil besitzen. Das vorliegende Design würde demzufolge die ermittelte längsschnittliche Veränderung positiv verzerren und daher scheint der Einsatz von Testdesigns wie in Tabelle 7.2 nicht empfehlenswert.

In einem zweiten Fall soll nun davon ausgegangen werden, dass sich die Itemblöcke A, B und C strukturell unterscheiden. Im Block A können daher „leichte“ Items enthalten sein, die bei einer Vorlage zu T2 „zu leicht“ sein könnten. Der Block B könnte für Schüler T1 schwer, zu T2 jedoch mittelschwer ausfallen. Im Block C könnten sich Items befinden, die Schüler zu T1 nicht beantworten können, weil sie sich auf Lerninhalte (oder Verfahren) beziehen, die erst im Zeitraum zwischen T1 und T2 unterrichtet wurden (*construct shift*; siehe Martineau, 2006). Alternativ könnte der Block C für Schüler zu T1 „zu schwierig“ sein. Unter diesem Gesichtspunkt trifft der Forscher eine bewusste Auswahl von Items als Ankeritems und das Testdesign. Da die Itemblöcke nicht austauschbar sind, bezieht man sich zu den Zeitpunkten auf (ggf. leicht) verschiedene Konstrukte. Nimmt man an, dass die Itemschwierigkeiten (als p-Werte) zu T1 und T2 im Mittel und bezüglich der Abfolge von Items nicht unterscheiden sollten, dann führt dies zwangsläufig zu verschiedenen Positionen des Blocks B zu T1 und T2. Ob demzufolge bei nicht austauschbaren Itemblöcken in Tabelle 7.2 die Veränderung unverzerrt abgebildet wird, ist eine Folge von mehreren Entscheidungen bezüglich des Testdesigns, die über die Frage einer konstant gehaltenen Itemposition der Ankeritems hinausgeht. Vorsichtig würde ich daher argumentieren, dass Testdesigns wie in Tabelle 7.2 auch zur adäquaten Beschreibung längsschnittlicher Veränderung geeignet sind.

Schwieriger ist die Frage der Definition der Fähigkeit im Hinblick auf interindividuelle Veränderung und der damit verbundenen Stabilität der Fähigkeiten zwischen T1 und T2. Formal können wir annehmen, dass zu den Zeitpunkten $t = 1$ (T1) und $t = 2$ (T2) die Fähigkeiten θ_{ptk} zu zwei Positionen $k = 1$ (erste Position) und $k = 2$ (zweite Position) in den Testheften existieren mögen. Das bedeutet, dass man zu T1 zwei Fähigkeiten θ_{p11} und θ_{p12} sowie zu T2 zwei Fähigkeiten θ_{p21} und θ_{p22} betrachtet. Daher kann ein vierdimensionales IRT-Modell für die Fähigkeiten θ_{ptk} postuliert werden. Falls ein Forscher die erste Testposition als Referenz wählt, so werden die Variablen θ_{p11} und θ_{p21} betrachtet, so liegt der Fokus der Analyse auf zwei der vier Dimensionen.

In einigen längsschnittlichen Untersuchungen werden für zwei Zeitpunkte strukturell ähnliche Testdesigns wie in Tabelle 7.1 eingesetzt. Dabei können für einige Personen Testhefte zu jeweils zwei Testpositionen vorliegen, für manche Personen liegen aber für ein oder beide Zeitpunkte nur Items zu einer Testposition vor. Daher gibt es einige Personen, für die keine Items an der ersten Testposition eingesetzt wurden. Wenn Fähigkeiten allerdings bezüglich der ersten Position definiert werden sollen, so müssen statistische Modellannahmen postuliert werden. Eine sinnvolle Inferenz für die erste Testposition scheint möglich, wenn man die gesamte Posteriorverteilung $\theta_{pk}^{(t)}$ für alle Zeitpunkte und alle Testpositionen simultan (etwa wiederum mit der Plausible Value Technik) auswertet. Liegen etwa zu T1 für eine Person nur Items an der zweiten Testposition vor, so nutzt man bei der Plausible Value Technik die aus anderen Testheften geschätzte Korrelation zwischen der ersten und der zweiten Testposition für die Generierung einer „hypothetischen“ Fähigkeit zur ersten Testposition für diese Person. Dabei sollten in entsprechenden Skalierungsmodellen (wie in (7.30) vorgeschlagen) zusätzlich bookletspezifische Positionseffekte eingesetzt werden, um Kontexteffekte adäquat abzubilden.

7.2 Bedeutung der Mehrebenenstruktur für IRT-Modelle

Zur Beantwortung empirischer Fragestellungen, für die Leistungstests administriert werden, enstehen häufig Datensätze mit Mehrebenenstrukturen (Snijders & Bosker, 2012). Die typisch anzutreffende Mehrebenenstruktur ist in Abbildung 7.3 dargestellt (vgl. Van den Noortgate et al., 2003). Formal handelt es sich dabei um eine kreuz-klassifizierte hierarchische Datenstruktur. Item Responses sind dabei einerseits in Personen und andererseits in Items (jeweils Level 2) geschachtelt. Personen sind wiederum in Klassen (oder Schulen) geschachtelt und Items sind in Itemgruppen (Testlets) geschachtelt (jeweils Level 3).

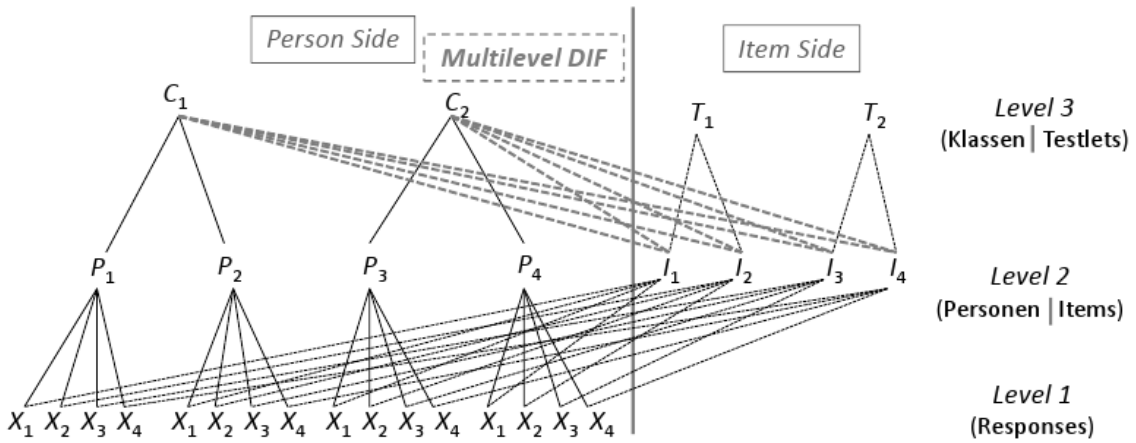

Abbildung 7.1: *Mehrebenenstruktur für Personen und Items*

In Abschnitt 3.4 und Kapitel 5 haben wir uns genauer mit der Mehrebenstruktur auf der Itemseite bei der Analyse von Testleteffekten und lokalen Abhängigkeiten beschäftigt. Auf die Mehrebenenstruktur auf der Personenseite sind wir in dieser Arbeit nur beiläufig eingegangen, weshalb wir dafür einige Anmerkungen in Abschnitt 7.2.1 treffen. Die Interaktion von Klassen und Items führt zu Analysen des Multilevel DIF, die in Abschnitt 3.5 vorgestellt wurden. Die Interaktion von Klassen und Testlets wurde spezifischer in Abschnitt 3.5.3 besprochen (siehe auch Jiao, Kamata, Wang & Jin, 2012 sowie Jiao & Zhang, 2015).

Im Folgenden diskutieren wir, welche Bedeutung die vorhandene Datenstruktur für die Modellwahl besitzt. Dabei muss unterschieden werden, ob die Mehrebenenstruktur explizit von substanziellem (analytischen) Interesse sind oder nur einen Störfaktor darstellen.

7.2.1 Mehrebenenstruktur als Störquelle

Sowohl für die Personenseite als auch die Itemseite kann die Mehrebenenstruktur in den Daten eine Störquelle (*nuissance factor*) darstellen. Wir gehen unter dieser Perspektive auf verschiedene Modellansätze ein.

Es wird davon ausgegangen, dass ein Leistungstest Schülern in Schulklassen vorgelegt wird. Dies führt zu einer Mehrebenenstruktur. In Large-Scale Assessments (wie z.B. PISA; OECD, 2014) sind Aussagen über die (totale) Population von Schülern von Interesse.

Da häufig ein komplexes zweistufiges stratifiziertes Stichprobendesign (OECD, 2014) zum Einsatz gelangt, werden statistische Replikationsmethoden wie Jackknife für eine valide statistische Inferenz eingesetzt (siehe z.B. Gershunskaya, Jiang & Lahiri, 2012 oder Kolenikov, 2010). In der Literatur dominieren Anwendungen und Empfehlungen, dass in Datensätzen mit Mehrebenenstrukturen Mehrebenenanalysen durchgeführt werden sollten, für alternative Standpunkte siehe jedoch Graubard und Korn (1994), Grömping (1996) oder VanderWeele (2010).

Wir gehen davon aus, dass (stetige Versionen der) Item Responses für Item i und Person p in Klasse s gemäß

$$X^*_{pis} = \theta_{ps} - b_{is} + \epsilon_{pis} \tag{7.31}$$

notiert werden können. In der totalen Population können die Fähigkeiten als ein Modell $\theta_{ps} = \mu + e_{ps}$ mit zentriertem normalverteiltem Fehler e_{ps} geschrieben werden. Der Mittelwert μ ist dann der Mittelwert über alle Schüler der Studie. Ein äquivalentes Mehrebenenmodell ist ein *random intercept model* Snijders und Bosker (2012) der Form $\theta_{ps} = \mu + u_s + r_{ps}$, das von einer Zerlegung in Klasseneffekte u_s und (Residual-)Effekten r_{ps} von Schülern *innerhalb* von Klassen ansetzt. Der Parameter μ ist nun allerdings nicht der Mittelwert über alle Schüler, sondern der Mittelwert über die (latenten) Mittelwerte der Fähigkeiten der Klassen s in dieser Studie[4]. Dies verdeutlicht, dass konzeptuell mit dem totalen Modell (Einebenenmodell) und dem Mehrebenenmodell verschiedene Parameter geschätzt werden. Häufig wird dies in der Diskussion zu praktischen Empfehlungen allerdings mit Effizienzfragen verwechselt[5]. ES sei betont, dass ein totales Modell auch bei substanzieller Intraklassenkorrelation eingesetzt werden kann, wenn die geschätzten Parameter aus diesem Modell von substanziellem Interesse sind.

Für die Skalierung (etwa mit dem Rasch-Modell) wird im Large-Scale Assessment häufig die Mehrebenenstruktur der Schachtelung von Schülern in Klassen nicht modelliert. Selbst wenn in (7.31) eine substanzielle Varianz von b_{is} existiert (also Multilevel DIF vorliegt), so ist man meist an einer Itemschwierigkeit für die totale Population interessiert. Das IRT-Modell (7.31) kann man dann umformulieren als

$$X^*_{pis} = \theta_{ps} - b_i + \underbrace{\{-\nu_{is} + \epsilon_{pis}\}}_{=\varepsilon_{pis}} \tag{7.32}$$

[4]Die Klassenmittelwerte v_s werden als $v_s = \mu + u_s$ definiert, folgen also der Verteilung $v_s \sim N(\mu, \sigma^2_B)$ mit $\sigma^2_B = Var(u_s)$. In der Schreibweise dieser hierarchischen Verteilung wird deutlich, dass μ der Mittelwert der (latenten) Klassenmittelwerte ist.

[5]Wenn das Mehrebenenmodell gilt (und damit das Sampling-Design nicht informativ ist), dann ist der Schätzer aus dem Mehrebenmodell effizienter als der OLS-Schätzer des Gesamtmittelwertes. Wenn das Mehrebenenmodell nicht gilt (und daher das totale Modell von analytischem Interesse ist), ist das Effizienzargument nicht mehr anwendbar und mit dem Mehrebenenmodell erhält man im Allgemeinen verzerrte Schätzungen. Manche Autoren argumentieren, dass auch für die Bestimmung von Standardfehlern nicht zwingend die Mehrebenenstruktur zu berücksichtigen sei (Rohwer & BloSSfeld, 2012). Dabei könnte man argumentieren, dass die statistische Inferenz nicht auf eine Population von Schülern abzielt, sondern auf ein datengenerierendes Modell (Kass, 2011). Beobachtete Kovariaten von Schülern werden dabei als Designvariablen interpretiert und Stochastizität entsteht nur durch einen Vorhersagefehler (bzw. ein Residuum) in einem statistischen Modell. Wenn die Annahme unabhängiger Residuen plausibel erscheint (und das ist in experimentellen und quasi-experimentellen Studien vielleicht der Fall), so müssen Standardfehler nicht adjustiert werden.

Anstelle des Fehlers ϵ_{pis} wird in (7.32) die Fehlerquelle $\varepsilon_{pis} = -\nu_{is} + \epsilon_{pis}$ betrachtet. Wie in Abschnitt 7.1.1 argumentiert, ist die Frage der wahren Traitvarianz $Var(\theta_{ps})$ und wahren Itemschwierigkeiten eine definitorische. Der Forscher muss entscheiden, ob die Interaktion von Items und Personen Bestandteil der (im IRT-Modell fixierten) Fehlervarianz sein soll oder nicht. Die Frage eines „Bias“ hat demnach nichts damit zu tun, ob bei der Anpassung eines Modells mit Multilevel DIF (7.31) substanzielle Varianzen $Var_s(b_{is})$ empirisch ermittelt werden.

Aus unserer Sicht sollte die Bestimmung der Standardfehler für Itemschwierigkeiten b_i das Sampling-Design berücksichtigen, was relativ einfach mit Jackknife anwendbar wäre (siehe Thomas & Cyr, 2002). Itemparameter werden *nicht* unabhängig von der Personenstichprobe geschätzt, weshalb die Cluster-Struktur für korrekte Standardfehler relevant ist.

In Kapitel 5 haben wir herausgearbeitet, dass auch die Clusterung von Items in Testlets nicht zwingend dazu führt, dass Modellparameter für Testlets (Testletvarianzen) von primärem Interesse sind. Die Bestimmung der „richtigen“ Reliabilität lässt sich dabei im Allgemeinen nicht empirisch beantworten. Selbst wenn behauptet wird, dass die Abhängigkeit von Items in einem gemeinsamen Testlet konstruktirrelavent ist, so ist die Spezifikation von Testlet-Modellen oder marginalen Modell (wie dem Copula-Modell in Abschnitt 5.3; Braeken, 2011) vor allem bei vielen Items schwierig umsetzbar. Daher bietet sich für Personen und Items ein vollständig designbasierter Ansatz (Binder & Roberts, 2012) an. Dabei wird das Rasch-Modell $X^*_{pi} = \theta_p - b_i + \epsilon_{pi}$ als Skalierungsmodell verwendet, dass keine Abhängigkeiten auf der Personenseite und der Itemseite berücksichtigt. Standardfehler für Itemschwierigkeiten können durch Jackknife der Personenstichprobe (etwa Jackknife von Klassen) erhalten werden. Die Inferenz für individuelle Fähigkeitswerte θ_p wird aber vom „Stichprobendesign“ der Items beeinflusst sein. Bei lokalen positiven Abhängigkeiten wird daher die statistische Inferenz auf Basis der Likelihood $\prod_i P(Y_{pi}|\theta)$ unter der Annahme lokaler stochastischer Unabhängigkeit zu unterschätzten Standardfehlern σ_{ind} führen (Ip, 2000). Wenn die Testlets, die die Abhängigkeit generieren, bekannt sind, so kann man einen adjustierten Standardfehler σ_{dep} für θ auf Basis einer modifizierten Likelihood (Ip, 2000) oder eines Ansatzes mit *generalized estimating equations* (GEE; Ip & Chen, 2012) berechnen. Das Verhältnis der beiden Standardfehler σ_{ind} und σ_{dep} kann genutzt werden, um korrekte Inferenz auf Basis der individuellen Posterior durchzuführen. Dabei wird die individuelle Posteriorverteilung um den Faktor $c = \sigma_{dep}/\sigma_{ind}$ gestreckt oder gestaucht und Plausible Values werden dann aus dieser modifizierten Posterior gezogen (siehe Bock, Brennan & Muraki, 2002 für ein ähnliches Vorgehen bei lokalen Abhängigkeiten für Ratingdaten).

Für Personenparameterschätzer (wie für den WLE; Warm, 1989) kann man individuelle Standardfehler auch durch Jackknife unter Weglassen von Items gewinnen[6]. Zur Berücksichtigung der Abhängigkeiten von Items in Testlets kann ein Jackknifing auf Basis von Testlets erfolgen (vgl. Monseur & Berezner, 2007 für Jackknifing von Testlets zur Berechnung von Linkingfehlern in PISA). Für eine korrekte Inferenz auf Basis der Posterior kann wiederum ein Faktor für die Adjustierung der Posterior (siehe oben) des Verhältnisses von Standardfehlern unter Berücksichtigung des komplexen Itemdesigns im

[6]Die Jackknife-Methode scheint gerade bei einer geringen Itemanzahl zu besseren Standardfehlern als die maximum likelihood basierte Schätzung zu kommen.

Verhältnis zum Standardfehler unter unabhängigen Items ermittelt werden. Auch stratifizierte Sampling-Verfahren (stratified Jackknife) können bei mehrdimensionalen Konstrukten (bzw. konstruktinhärenten Abhängigkeiten) oder bei finiten Itemstichproben eingesetzt werden. Die auf die Itemseite übertragenenen Jackknife-Verfahren scheinen häufig einfacher spezifizierbar zu sein als IRT-Modelle, die die Abhängigkeiten direkt modellieren.

7.2.2 Mehrebenenstruktur von substanziellem Interesse

Für die Klassifikation von Items und Klassen präsentieren Van den Noortgate und De Boeck (2005) eine Klassifikation als fixed bzw. random effects, die in Tabelle 7.4 schematisch dargestellt ist. Dabei wird der Status der klassenspezifischen Itemschwierigkeiten b_{is} festgelegt.

Tabelle 7.4: *Items und Klassen als fixed bzw. random effects*

	Klassen s	
Items i	Fixed	Random
Fixed	(a)	(b)
Random	(c)	(d)

Im Fall (a) werden Items und Klassen jeweils als fixed angesehen. Eine Analyse der Itemschwierigkeiten entspricht dem üblichen Vorgehen der Untersuchung differenziellen Itemfunktionierens bzw. der Invarianztestung bei mehreren Gruppen (Millsap, 2011). Der Fall (b) entspricht dem Multilevel DIF (*ML-DIF*) Ansatz. Für jedes feste Item werden Klassen als zufällig betrachtet und die Varianz der Klasseneffekte wird als Multilevel DIF Varianz bezeichnet (siehe Abschnitt 3.5). Für jedes Item wird dabei unabhängig eine eigene Multilevel DIF Varianz bestimmt. Im Fall (c) betrachtet man die Items als zufällig und die Klassen als fest. Im Fall zweier Gruppen wird dabei die Varianz der Differenzen der Itemschwierigkeiten als *DIF Varianz* bezeichnet (Longford, Holland & Thayer, 1993). Im Fall (d) werden sowohl Items als auch Klassen als zufällig betrachtet. Die Interaktion beider Effekte führt demzufolge zu einer (homogen geschätzten) Varianzquelle. Im Rahmen der Generalisierbarkeitstheorie ist diese Varianzquelle für die Berechnung von Fehlern bei Vergleichen von Klassenmittelwerten relevant (Kane, Gillmore & Crooks, 1976; Brennan, 2001a, S. 130ff.). Die Schüler innerhalb von Klassen werden immer als zufällig betrachtet.

In der stochastischen Messtheorie (Steyer & Eid, 2001) werden Messfehler nur intraindivuell definiert. Demzufolge können im Rahmen dieser Theorie Fehler nur auf individueller Ebene entstehen und daher sind im Rahmen dieser Theorie keine Interaktionen von Items und Klassen zulässig, sondern verstoßen gegen das Messmodell (vgl. auch Eid & Koch, 2014). Jak, Oort und Dolan (2013) schlagen einen sog. Test auf *cluster bias* vor, mit dem simultan getestet wird, ob die Varianz aller klassenspezifischen Itemschwierigkeiten gleich Null ist und empfehlen den Einsatz des Tests bei Vorliegen von Mehrebenendaten bei der Betrachtung von Individualkonstrukten. Für Leistungstests argumentieren wir, dass die Existenz von Multilevel DIF nicht zwingend einen „Bias“ darstellt, sondern vielmehr eine gewünschte Variationsquelle ist (siehe den nachfolgenden Abschnitt 7.2.3).

Das Multilevel DIF Modell kann in der Form

$$X^*_{pis} = u_s + r_{ps} - b_i - \nu_{is} + \epsilon_{pis} \tag{7.33}$$

mit klassenspezifischen Itemschwierigkeiten $b_{is} = b_i + \nu_{is}$ und Fähigkeiten $\theta_{ps} = u_s + r_{ps}$ notiert werden. Fox (2010, Kap. 7) diskutiert sog. random item models (siehe auch Fox & Verhagen, 2010), in denen gegenüber (7.33) auch klassenspezifische Itemladungen $\lambda_{is} = \lambda_i + \alpha_{is}$ eingeführt werden

$$X^*_{pis} = (\lambda_i + \alpha_{is})(u_s + r_{ps}) - b_i - \nu_{is} + \epsilon_{pis} \tag{7.34}$$

Durch Einführung der zufälligen Effekte α_{is} ist nun auch eine Variation der Itemladungen zwischen Klassen zugelassen. Multipliziert man die Gleichung (7.34) aus, so ergibt sich

$$X^*_{pis} = \lambda_i u_s + \lambda_i r_{ps} - b_i + \underbrace{\alpha_{is} u_s - \nu_{is}}_{\text{Level 2}} + \underbrace{\alpha_{is} r_{ps} + \epsilon_{pis}}_{\text{Level 1}} \tag{7.35}$$

Gleichung (7.35) zeigt, dass die klassenspezifische Variation der Ladungen mit der Variation in den Intercepts ist, wenn die Varianz auf den Ladungen unspezifiziert bleibt. Ein Teil der Varianz ($\alpha_{is} r_{ps}$) ist mit der Fehlervarianz konfundiert. Wenn Beurteilungen der Leistungstests auf Klassenebene stattfinden (wie in deutschen Lernstandserhebungen), so könnten bestimmte Items klassenspezifisch niedrige Itemladungen aufweisen (vgl. Spoden, Fleischer & Leutner, 2014). Häufig werden auch Zweiebenen-Faktorenanalysen zur Analyse der Struktur des Leistungstests eingesetzt (Goldstein et al., 2007). Dabei erhält man neben der klassenspezifischen Itemvarianz ν_{is} (in (7.35) eine Itemladung λ_i^B auf der Ebene der Klassen und eine Itemladung λ_i^W auf der Ebene der Schüler innerhalb von Klassen. Wenn diese beiden Ladungen (bei geeigneter Normierung) verschieden ausfallen, so ist dies weder notwendig noch hinreichend dafür, dass klassenspezifische Variation in Itemladungen – wie in (7.34) definiert – auftritt[7].

Wahl der Linkfunktion in Mehrebenenmodellen mit dichotomen Daten

In Gleichung (7.33) werden klassenspezifische Itemschwierigkeiten $b_{is} = b_i + \nu_{is}$ angenommen. Spezifiziert man stattdessen ein Rasch-Modell, so geht bei Gültigkeit von (7.33) die Varianzquelle $Var(\nu_{is})$ in die fixierte Fehlervarianz (neben ϵ_{pis}) über. Dies verdeutlicht mit Hilfe der Überlegungen in Abschnitt 7.1.1, dass im ML-DIF-Modell (7.33) die extrahierte Traitvarianz um den Faktor $(E_i\left[Var_s(\nu_{is}^2)\right]/D^2 + 1)$ höher ausfällt[8]. Die aus dem ML-DIF-Modell ermittelten mittleren Itemschwierigkeiten b_i sind demzufolge nicht auf derselben Metrik wie im Rasch-Modell. Diese Eigenschaft des logistischen Modells besitzt die Spezifikation des linearen Modells (*linear probability model*; Angrist & Pischke, 2008) nicht. Für die dichotomen Item Responses X_{pis} wird ein „gewöhnliches“ lineares Modell mit gemischten Effekten angepasst

$$X_{pis} = u_s + r_{ps} - b_i - \nu_{is} + \epsilon_{pis} \tag{7.36}$$

[7] In Jak et al. (2013) wird implizit formuliert, dass eine Äquivalenz beider Modellansätze gegeben sei. In späteren Arbeiten der Autoren ist dieses Argument allerdings nicht mehr zu finden (Jak, Oort & Dolan, 2014).

[8] Mit dem Symbol $E_i\left[Var_s(\nu_{is}^2)\right]$ ist der Mittelwert der ML-DIF-Effekte über alle Items gemeint.

Bei einem Hinzufügen klassenspezifischer Effekte bleibt in diesem Modell $Var(X_{pis})$ konstant, weshalb die Größe der Varianzkomponenten über verschiedene Modelle hinweg vergleichbar bleibt. Interessanterweise wird in der Mehrebenenliteratur vorgeschlagen, die Varianzaufklärung für logistische Mehrebenenmodelle in der originalen Metrik der dichotomen Variablen (also mit dem Ansatz (7.36)) durchzuführen (Goldstein, 2011, S. 127ff.). Die Wahl des linearen Modells anstelle des logistischen Modells wird in der ökonometrischen Literatur mitunter kontrovers diskutiert. In der Item Response Theorie wird häufig der Nachteil betont, dass die Item-Response-Funktionen im linearen Modell vorhergesagte Werte (Wahrscheinlichkeiten) kleiner als Null oder größer als Eins sein können (McDonald, 1999; Rost, 2004). Diesen Kritikpunkt teile ich im Allgemeinen nicht, im Gegenteil: mit dem logistischen Modell werden die vorgehersagten Werte im Modell tendenziell verzerrt sein (siehe Freedman et al., 2008). Wenn man mit dem logistischen Modell Vorhersagen $P(Y|X = x_p)$ für eine individuelle Kovariate X treffen will und die Wahrscheinlichkeit $P(Y|X)$ in der Population nahe bei Eins liegt, so wird mittels der logistischen Regression sicher eine verzerrte Schätzung erhalten, da in der logistischen Regression nur Werte kleiner als Eins vorhergesagt werden können. Die Eigenschaft der Erwartungstreue in der logistischen Regression bezieht sich also auf die Logitmetrik, nicht die Originalmetrik der Wahrscheinlichkeiten.

Maydeu-Olivares (2005) vergleicht die Anpassungsgüte eines linearen Modells mit einem IRT-Modell für den häufig verwendeten Beispieldatensatz LSAT6 und findet, dass das lineare Modell praktisch einen gleich guten Fit (gemessen in vorhergesagten bivariaten Korrelationen zwischen Items) aufweist.

Letztendlich scheint aber die Wahl der Metrik für die Erfassung der Variation von Itemschwierigkeiten in einem gewissen Sinne willkürlich (Goldstein, 1980). Der Forscher kann definieren, ob Variabilität in Lösungsverhalten auf den Items in der Metrik der Lösungshäufigkeiten (der p-Werte und damit dem linearen Modell) oder dem logistischen Modell geschehen soll. Ein Verweis, dass für Wahrscheinlichkeiten in einem mittleren Bereich (etwa zwischen .2 und .8) die lineare und die logistische Linkfunktion praktisch zu denselben Implikationen bezügliche Itemparametern und geschätzter Traitvarianz führen, könnte bei der konkreten Modellauswahl für beide Linkfunktionen als Argument für die Verwendung gewählt werden[9].

Tabelle 7.5: *Multilevel DIF für drei Items und drei Klassen*

	lineares Modell					logistisches Modell				
Item	M	SD	Kl1	Kl2	Kl3	M	SD	Kl1	Kl2	Kl3
I1	.47	.08	.40	.55	.47	-.11	.30	-.41	.20	-.12
I2	.81	.08	.81	.73	.88	1.48	.50	1.45	.99	1.99
I3	.47	.12	.35	.59	.47	-.13	.49	-.62	.36	-.12

Anmerkung: Kl1, Kl2, Kl3 bezeichnen drei Klassen.

In Tabelle 7.5 sind beispielsweise die Konsequenzen der Wahl der Linkfunktion für drei Items und drei Klassen dargestellt. Die drei Klassen sollen dabei illustrativ die Po-

[9]In diesem Fall wäre das computationale weniger aufwändige Modell – das lineare Modell – vorzuziehen.

pulation der Klassen darstellen, für deren Itemschwierigkeiten wir Mittelwerte (M) und Standardabweichungen (SD) berichten. Im linearen Modell werden dabei p-Werte, im logistischen Modell $logit(p)$ verwendet. Im linearen Modell haben dabei die Items I1 und I2 die gleiche Standardabweichung von .08, aber verschiedene Itemschwierigkeiten. Wählt man die logistische Transformation der Wahrscheinlichkeiten, dann ist die Variabilität der Itemschwierigkeiten für das leichtere Item I2 auf der Logitmetrik (SD=.50) höher als für Item I1 (SD=.30). Dies liegt darin begründet, dass die logistische Transformation extreme Wahrscheinlichkeiten stärker spreizt als mittlere Wahrscheinlichkeiten. Die Items I2 und I3 haben näherungsweise die gleiche ML-DIF-Varianz auf der Logitmetrik (SD=.50 bzw. SD=.49). Auf der Metrik der Wahrscheinlichkeiten im linearen Modell besitzt dann aber das Item I3 mit der mittleren Schwierigkeit eine höhere Varianz als das leichtere Item I2. In Abhängigkeit der Wahl der Linkfunktion unterscheiden sich also Aussagen darüber, welche Items am stärksten zwischen Klassen variieren. Mitunter wird argumentiert, dass man „bessere Schätzungen“ des Multilevel DIF bei möglichen „Bodeneffekten“ bzw. „Deckeneffekten“ der Itemschwierigkeiten durch Wahl der logistischen anstelle der linearen Linkfunktion erhält (vgl. Embretson & Reise, 2000, S. 34ff.; Embretson & Poggio, 2012). Dabei fußt die Argumentation auf der Annahme, dass das logistische Modell das datengenerierende Modell ist und man bei einer Anpassung des „falschen“ linearen anstelle des logistischen Modells zu verzerrten Schätzungen gelangt. Geht man umgekehrt davon aus, dass das lineare Modell das datengenerierende Modell ist und würde das logistische Modell anpassen, so „zeigt“ man, dass das logistische Modell zu verzerrten Schätzern führt. Auf Basis von Simulationsstudien ist daher die Modellwahl nicht begründbar.

Aus einer Modellfit-Perspektive wäre das lineare oder das logistische Modell nur eines von vielen anderen möglichen Linkfunktionen. Man könnte die Linkfunktion g so schätzen, dass der Modellfit maximiert wird, d.h. man schätzt eine Linkfunktion g, so dass

$$P(X_{pis} = 1) = g\left(u_s + r_{ps} - b_i - \nu_{is}\right) \tag{7.37}$$

Die Linkfunktionen g sind dabei als Funktionsklasse parametrisiert, d.h. $g = g_{\boldsymbol{\psi}}$ mit einem zu schätzenden Parameter $\boldsymbol{\psi}$ (vgl. Stukel, 1988; Koenker & Yoon, 2009). Die Linkfunktionen werden dabei typischerweise monoton gewählt, können aber asymmetrisch sein. Itemspezifische asymmetrische Linkfunktionen sind auch für IRT-Modelle vorgeschlagen worden (Bazán, Branco & Bolfarine, 2006; siehe auch Samejima, 1997, 2010 für Reviews).

7.2.3 Multilevel DIF: Fähigkeiten als Level-2-Konstrukt

In Abschnitt 7.2.1 wurde argumentiert, dass Multilevel DIF als Interaktion von Items und Klassen in nationalen und internationalen Vergleichsstudien (wie in PISA) eher als Störquelle und daher nicht von substanziellem Interesse ist. Für Leistungstests, die Rückmeldungen an Klassen (und Schulen) liefern (wie Vergleichsarbeiten, VERA, im Folgenden integrierend als *Lernstandstests* bezeichnet; siehe Hosenfeld, 2008), könnte diese Einschätzung jedoch anders ausfallen. Wenn Rückmeldungen bei Lernstandstests „Lernerträge“ des unterrichtlichen Geschehens abbilden sollen, so könnte die zu erfassende Fähigkeit auf der Ebene der Klassen definiert und daher ein *Level-2-Konstrukt* (Lüdtke, Robitzsch, Trautwein & Kunter, 2009; Zumbo & Forer, 2011) sein. Im Hinblick auf die Itemauswahl sollte dabei primär Variation zwischen Klassen und weniger zwischen Schüler generiert werden,

da Fähigkeiten von Klassen klassifiziert werden sollen. Werden Items gemäß einer maximalen internen Konsistenz ausgewählt, so stehen bei Level-2-Konstrukten die Itemladungen λ_i^B auf der Klassenebene (Level 2) im Rahmen einer konfirmatorischen Faktorenanalyse (siehe Goldstein et al., 2007) im Vordergrund. Die psychometrischen Eigenschaften der Items auf Individualebene (Level 1) würden dann weniger (oder keine) Relevanz erlangen. Eine Erhöhung der internen Konsistenz in der totalen Population und daher eine Auswahl der Items mit hohen Itemladungen wird im Allgemeinen dazu führen, Items mit großem Multilevel DIF zu entfernen. Zur Verdeutlichung dieser Aussage nehmen wir an, dass ein Multilevel DIF Modell mit Itemladungen gelten möge

$$X_{pis}^* = a_i\theta_{ps} - b_i - \nu_{is} + \epsilon_{pis} \tag{7.38}$$

Hierbei bezeichnet a_i die Itemladung (gleich gesetzt zwischen Level 2 und Level 1) bei gleichzeitigem Zulassen von Multilevel DIF durch zufällige Effekte ν_{is}. Passt man nun allerdings ein 2PL-Modell unter Ignorierung des Multilevel DIF an, so sind nach den Überlegungen in Abschnitt 7.1.1 die Varianzquellen ν_{is} und ϵ_{pis} konfundiert. Da die Residualvarianz im logistischen Modell fixiert wird, beträgt die ermittelte Itemladung a_i^* im 2PL-Modell

$$a_i^* = \frac{a_i}{\sqrt{Var(\nu_{is})/D^2 + 1}} \tag{7.39}$$

Die Trennschärfe a_i^* wird demzufolge deutlich kleiner als a_i, falls die Multilevel DIF Varianz $Var(\nu_{is})$ groß ist.

Gilt das Modell der eindimensionalen Mehrebenen-Faktorenanalyse, so ist

$$X_{pis}^* = \lambda_i^B u_s + \lambda_i^W r_{ps} - b_i - \nu_{is} + \epsilon_{pis} \tag{7.40}$$

Wir nehmen dabei an, dass die Ladungen geeignet standardisiert sind, z.B. durch die Bedingung $\prod_i \lambda_i^B = \prod_i \lambda_i^W = 1$. Passt man nun allerdings ein 2PL-Modell an und nutzt die resultierenden Itemladungen a_i^* als Itemselektionskriterium, so folgt (bei nichtinformativem Cluster-Design)

$$a_i^* = \frac{\lambda_i^B \rho_I + \lambda_i^W (1 - \rho_I)}{\sqrt{Var(\nu_{is})/D^2 + 1}} \tag{7.41}$$

wobei ρ_I die Intraklassenkorrelation von θ bezeichnet. Die Trennschärfe a_i^* setzt sich demzufolge aus der Itemladung auf Klassenebene und der Itemladung auf Individualebene zusammen. Bei (typischerweise) nicht all zu hohen Intraklassenkorrelationen (z.B. $\rho_I < .30$) dominiert dann die Itemladung der Individualebene die Itemauswahl.

Im Fall von Faktormodellen mit stetigen Items diskutieren Pornprasertmanit, Lee und Preacher (2014) die Konsequenzen einer Vernachlässigung der Mehrebenenstruktur im Hinblick auf Itemparameter und Kovarianzstrukturen. Sie leiten dabei ähnliche Formeln wie in (7.41) ab.

Zusammenfassend sollen die Betrachtungen zeigen, dass übliche an der totalen Population von Schülern orientierte psychometrische Kriterien für die Itemauswahl bei Lernstandstests nicht zwingend geeignet sind, da das interessierende Merkmal primär die Ebene der Klassen (Level 2) sein könnte. Die empirische Relevanz unserer Überlegungen bleibt natürlich für konkrete Anwendungen nachzuweisen.

7.2.4 Multilevel DIF und Instruktionssensivität

Klassenspezifische Itemschwierigkeiten können auf klassenspezifische Lerngelegenheiten (*opportunity to learn*, OTL; Muthén et al., 1991) hinweisen. Die Intensitität des OTL wird dabei in der Varianzquelle des Multilevel DIF abgebildet. In Abschnitt 3.5.2 wurde Multilevel DIF für den DEMAT4 (Gölitz et al., 2006) untersucht. Im zugehörigen DEMAT4-Manual (Gölitz et al., 2006) sind Lösungshäufigkeiten für die Items in der Mitte der vierten Klassenstufe und zum Ende der vierten Klassenstufe in einer Normstichprobe aufgeführt. Da der Testeinsatz für DEMAT4 im zweiten Halbjahr der vierten Klassenstufe vorgesehen ist, kann man die Lösungshäufigkeiten als Prätest und Posttest zwischen der Instruktionsphase ansehen. In der Literatur bezeichnet man die Differenz aus Lösungshäufigkeiten des Posttests und des Prätests als *instructional sensitivity* (Polikoff, 2010). Für den DEMAT4 zeigt sich, dass die logarithmierten Multilevel DIF Varianzen mit der Differenz der Lösungshäufigkeiten zu $r = .57$ korrelieren (Robitzsch, 2011b). Querschnittlicher Multilevel DIF hängt demzufolge mit der Variation in längsschnittlicher Veränderung zusammen. Items mit hohem Multilevel DIF könnten daher besonders valide sein, da sie besonders instruktionssensitiv sind (vgl. auch Briggs, 2011).

Naumann, Hochweber und Hartig (2014) schlagen Multilevel DIF als Maß instruktionaler Sensitivität für Längsschnittdaten vor. Für eine Person p in Klasse s, Item i und Zeitpunkt t nehmen sie folgendes IRT-Modell an

$$X^*_{pist} = \theta_{pst} - b_{ist} + \epsilon_{pist} \tag{7.42}$$

Die Itemschwierigkeiten $b_{ist} = b_{it} + \nu_{ist}$ werden dabei in zeitpunktspezifische feste Effekte und klassenspezifische zufällige Effekte dekomponiert. Als instruktionale Sensitivität wird dann die Differenz der Itemschwierigkeiten zwischen Posttest ($t = 2$) und Prätest ($t = 1$) definiert:

$$\Delta b_{is} = b_{it2} - b_{it1} = b_{i2} - b_{i1} + \nu_{is2} - \nu_{is1} = \mu_i + \Delta\nu_{is} \tag{7.43}$$

Der Parameter μ_i bezeichnet dabei als Posttest-Prätest-Differenz (PPD) die *globale Sensitivität*, während die Varianz $\phi_i = Var(\Delta\nu_{is})$ als *differenzielle Sensitivität* (auch PPD Varianz) bezeichnet wird. Das vorgeschlagene IRT-Modell (7.42) erlaubt damit eine itemspezifische Quantifikation von globaler und differenzieller Sensitivität und verallgemeinert daher das in Abschnitt 3.5 vorgeschlagene Multilevel DIF Modell.

Gerade bei vielen Items wird man für eine leichtere Interpretierbarkeit das Modell (7.42) um weitere hierarchische Varianzkomponenten erweitern. Die Varianz $Var(\mu_i)$ kann dabei die Variation in der globalen Sensitivität kennzeichnen und ist damit auch eine Effektgröße längsschnittlichen differenziellen Itemfunktionierens (siehe auch Kapitel 4 dieser Arbeit). Die Variabilität der differenziellen Sensitivität zwischen Items ließe sich mit $Var(\log \phi_i)$ erfassen. Diese Betrachtung entspricht einem Modell der Generalisierbarkeitstheorie (Brennan, 2001a) in Erweiterung auf die logistische Linkfunktion (vgl. Glas, 2012b), in denen einzelne Einheiten spezifische Varianzen besitzen und zusätzlich hierarchische Verteilungen für diese spezifischen Varianz angenommen werden (siehe auch Hedeker, Mermelstein & Demirtas, 2008 für das verwandte *mixed effects location scale model*).

Vor allem bei der Erfassung der Variabilität von Veränderungen ist die Wahl der Metrik bedeutsam. Im Abschnitt 7.2.2 haben wir den Einsatz der linearen (identischen) gegen-

über der logistischen Linkfunktion abgewogen. Für den verwendeten IGEL-Datensatz in Naumann et al. (2014, S. 389ff.) erkennt man, dass die Items zum Prätest im Mittel relativ schwierig sind. Betrachtet man nun längsschnittliche Veränderung für schwierige und mittelschweren Items, so wird die Veränderung schwieriger Items in der Logitmetrik gegenüber mittelschweren Items typischerweise höher ausfallen. Ob dies „die Realität" geeignet abbildet, ist nur im Lichte der Betrachtung der Validität der in Naumann et al. (2014) getroffenen Aussagen zu beurteilen (siehe wiederum auch Kapitel 4).

7.3 Rolle des Domain Samplings in der Item-Response-Theorie

In diesem Abschnitt diskutieren wir IRT-Modelle primär unter der Perspektive, dass wir von einer Population von Items ausgehen, aus der für einen konkreten Test nur endlich viele Items ausgewählt werden. Wir führen zuerst das Konzept des Domain Samplings in Abschnitt 7.3.1 ein und stellen Bezüge zur Generalisierbarkeitstheorie in Abschnitt 7.3.2 her. Anschließend wird die Domain Sampling Perspektive für die Rechtfertigung für Cronbachs Alpha in Abschnitt 7.3.3 näher beleuchtet. Die Bedeutung von Domain Sampling und Faktormodellen diskutieren wir in Abschnitt 7.3.4. Abschließend gehen wir auf die Bedeutung des Domain Samplings für Invarianztestungen in Abschnitt 7.3.5 ein.

7.3.1 Domain Sampling

Im Ansatz des *Domain Samplings* (bzw. Item Sampling; Nunnally & Bernstein, 1994; Revelle, 2015, Kap. 7) wird davon ausgegangen, dass die in einem Test eingesetzten Items eine Domäne (Population) von Items repräsentieren (siehe auch Abschnitt 4.3.1). In Analogie zum Sampling von konkreten Personen aus einer Population von Personen inferieren statistische Modelle für einen konkreten Test mit endlich vielen Items auf die Population aller Items. Abstrakt gesprochen kann daher studiert werden, welche Eigenschaften ein den Test charakterisierender mehrdimensionaler Parameter $\boldsymbol{\gamma}_I$ (zum Beispiel eine mittlere Kovarianz zwischen Items) im Hinblick auf den Parameter $\boldsymbol{\gamma}_\infty$ besitzt, der die (meistens) unendliche Population von Items beschreibt.

Unter Sampling von Items muss dabei keine konkrete zufällige Auswahl von Items verstanden werden. Vielmehr werden statistische Modelle zur Beschreibung des Samplings eingesetzt, die annehmen, dass die verwendete Itemstichprobe im Test repräsentativ für die Itempopulation ist. Damit unterscheidet sich diese Überlegung nicht vom Sampling von Personen in wissenschaftlichen Untersuchungen. Oftmals erfolgt das Sampling von Personen auch nicht zufällig und der Einsatz statistischer Modelle erfolgt „nur" unter der Annahme, dass die Stichprobe von Personen repräsentativ für die Population von Personen ist.

In Abbildung 7.2 ist das Sampling von Items schematisch dargestellt. In einer Testsituation erfolgt dabei das Sampling häufig stratifiziert, symbolisiert in der Abbildung durch drei Subpopulation, die man als Testkomponenten bezeichnen kann. Diese Testkomponenten gehen mit verschiedener Gewichtung in einen Skalenwert θ ein, damit Repräsentativität gegeben ist. Beispielsweise kann die Administration von Items einer Testkom-

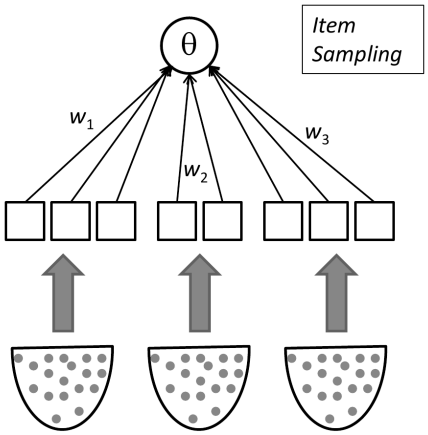

Abbildung 7.2: *Schematische Darstellung des stratifizierten Item Sampling*

ponente mit großem ökonomischen Aufwand verbunden sein, weshalb nur wenige Items dieser Testkomponente im Test vorgelegt werden. Dann müssen die Items der jeweilige Testkomponenten hochgewichtet werden, um den „richtigen“ Schätzer der Fähigkeit θ zu erhalten.

Für einen eingesetzten Test mit I Items, die durch Domain Sampling entstehen, kann man sich als Schätzer einer Personenfähigkeit θ_p eine Vorschrift $\theta_{pI} = 1/I \cdot \sum_i a_i X_{pi}$ vorstellen[10]. Die Personenfähigkeit in der Population aller Items ist dann definiert als Grenzwert der Fähigkeit für Tests mit wachsender Testlänge I, d.h.

$$\theta := \lim_{I \to \infty} \theta_{pI} \tag{7.44}$$

Die latente Variable θ, die die Fähigkeit beschreibt, hängt also definitorisch an einer unendlichen Itempopulation (vgl. auch Stout, 1990; Junker, 1991; Ellis & Junker, 1997). Ein perfekt reliabler Test würde sich demzufolge ergeben, wenn alle Items der Population im Test eingesetzt werden. In der gerade zitierten Literatur werden Bedingungen diskutiert, die die Existenz einer *essenziell eindimensionalen* latenten Variablen θ mit monoton wachsenden Item-Response-Funktionen sichern (siehe Stout, 1990). Es zeigt sich dabei, dass man für die Existenz eines solchen θ schwächere Konzepte als die lokale stochastische Unabhängigkeit benötigt (Stout, 1990). Beispielsweise kann eine essenziell eindimensionale latente Variable θ auch in einem Test mit Testlets mit nicht verschwindenden positiven lokalen Abhängigkeiten definiert sein, da der Einfluss einzelner Testlets in einem unendlich langen Test verschwindet (vgl. auch Kapitel 5 dieser Arbeit)[11].

Im Domain Sampling wird daher die latente Variable primär durch das Argument legitimiert, dass die Items im Test repräsentativ für die Domäne sind. Damit ist dieses Konzept nahe an der Idee formativer Messmodelle, in denen Items die latente Variable θ definieren, also wie in Abbildung 7.2 Veränderungen von Itemausprägungen Veränderungen in θ nach sich ziehen und nicht umgekehrt, wie dies in reflektiven Messmodellen

[10] Die Itemladungen a_i mögen dabei geeignet normiert sein, d.h. $1/I \cdot \sum_{i=1}^{I} a_i \to 1$ für $I \to \infty$.

[11] Hierbei muss man nur annehmen, dass die Menge der Items in einem Testlet beschränkt sein soll, d.h. eine bestimmte Anzahl von Items(zum Beispiel zehn Items) darf nicht überschritten werden.

der Fall ist. In Abschnitt 3.2 wurde diskutiert, dass das Rasch-Modell als reflektives oder als formatives Messmodell interpretiert werden. Markus und Borsboom (2013a, 2013b) argumentieren, dass man das Modell des Domain Samplings (für unendlich viele Items) ebenso als reflektives Messmodell verstehen sollte. Dabei verwenden sie den Begriff einer homogenen Itemdomäne (McDonald, 2003), den sie als *vanishing conditional independence* (VCI) im Sinne von Ellis und Junker (1997) interpretieren. Letztgenannte Eigenschaft ist praktisch die lokale stochastische Unabhängigkeit von Items, bei der allerdings nicht auf θ, sondern auf die Itemausprägungen im Unendlichen $(X_n, X_{n+1}, \ldots)$ bedingt wird. Ich argumentiere, dass das Konzept des Domain Samplings bei der Existenz mehrerer Testkomponenten im Allgemeinen nicht zu einer eindimensionalen latenten Variablen führen wird, da die Annahme des VCI zu strikt scheint (siehe auch einen entsprechenden Kommentar in Ellis & Junker, 1997), weil es ein Domain Sampling mit mehreren Itemstrata (die wir auch mit Testkomponenten oder Dimensionen assoziieren) nicht einschließt.

Unendliche Itempopulationen werden ebenso im Kontext von explorativen und konfirmatorischen Faktorenanaysen diskutiert (Kaiser & Caffrey, 1965; Levine & Hunter, 1971; McDonald, 1978; McDonald & Mulaik, 1979; Mulaik, 2009a; Williams, 1978, 1979). Dabei wird das Konzept der statistischen Inferenz (Inferenz auf eine Population von Personen) von der psychometrischen Inferenz (Inferenz auf eine Population von Items) unterschieden (Husek & Sirotnik, 1967). Demzufolge können die den Test beschreibenden Zufallsvariablen der Items durch ein Sampling in den beiden Facetten Personen und Items beschrieben werden. Hunter (1968) schägt für den Kontext der Generalisierbarkeitstheorie (Cronbach et al., 1972) eine geeignete maßtheoretische Formalisierung vor. Ein auf Resampling-Verfahren basierende Inferenz für beide Facetten kann das sog. *Double Jackknife* darstellen, bei dem einzelne Personen oder Personengruppen bzw. Item oder Itemgruppen für Analysen entfernt werden (vgl. Brennan, 2001a, S. 182ff.; Haberman, Lee & Qian, 2009; Xu & von Davier, 2010). Die Methode des Double Jackknife ist dabei für beliebige statistische Modelle einsetzbar. Im Kontext der Homogenitätsanalyse als Verfahren der dimensionalen Analyse kategorialer Variablen diskutieren Michailidis und de Leeuw (1998) die *Stabilität* von Parameterschätzungen unter Weglassen einzelner Personen oder Personengruppen bzw. von Items.

Als Beispiel für eine gleichzeitige statistische und psychometrische Inferenz (die sog. *statistisch-psychometrische Inferenz* nach Husek & Sirotnik, 1967) kann man sich das Reliabilitätsmaß Cronbachs Alpha (α) vorstellen (siehe Abschnitt 7.3.3). Statistische Unsicherheit in α entsteht dabei durch das Sampling von Personen und durch ein Sampling von Items, wenn man davon ausgeht, dass die eingesetzten Items im Test nicht die komplette Domäne darstellen. Die Reliabilität α kann bei einem Item Sampling selbst bei sehr großen Personenstichproben demzufolge einen bedeutsamen „Standardfehler“ aufweisen.

7.3.2 Item-Response-Theorie und Generalisierbarkeitstheorie

Die Generalisierbarkeitstheorie (G-Theorie; Cronbach et al., 1972; Brennan, 2001a) beschreibt das Sampling mehrere Facetten. Diese können beispielsweise Personen, Items, Rater, Situationen oder Zeitpunkte sein. Als statistisches Modell dient ein Varianzkomponentenmodell (Searle et al., 1992). Im Spezialfall des Domain Samplings werden Items und Personen gesampelt. Bei vorliegenden N Items und I Personen entsteht eine $N \times I$-

Matrix $\boldsymbol{X} = (X_{pi})_{pi}$ von Beobachtungen von Person p auf Item i. Die korrespondierende Matrix in den Populationen von Personen und Items entsteht, in dem man sowohl Zeilen als auch die Spalten in $\boldsymbol{X}$ gegen Unendlich gehen lässt (Cronbach & Shavelson, 2004). Das statistische Modell der Generalisierbarkeitsheorie mit den Facetten Personen und Items ist

$$X_{pi} = \mu + \nu_p + \nu_i + e_{pi} \tag{7.45}$$

Dabei nimmt man $Var(\nu_p) = \sigma_p^2$ und $Var(e_{pi}) = \sigma_e^2$ an. Die Itemeffekte können entweder als fest (kohärent zu Faktorenanalysen und Item-Response-Modellen) oder als zufällig mit $Var(\nu_i) = \sigma_i^2$ angenommen werden. In der Schreibweise von Varianzkomponenten geht (7.45) über in

$$Var(X_{pi}) = \mu + \sigma_p^2 + \sigma_i^2 + \sigma_e^2 \tag{7.46}$$

Die Schätzung der Größen in (7.45) und (7.46) bezeichnet man auch als *G Theory*. Mit Hilfe der Varianzkomponenten kann beispielsweise für den Personenmittelwert $\overline{X}_{p\bullet}$ eine Fehlervarianz σ_e^2/I bestimmt werden, wenn man die Varianzquelle der Items als fixiert (und für relative Personenreihenfolgen) und als nicht relevant ansieht. Die *D Theory* trifft Aussagen über das Design zu planender Studien, in denen für interessierende Parameter vorgegebene Messfehlervarianzen eingehalten werden sollen (siehe Brennan, 2001a).

Aus einer Anwendungsperspektive wird man häufig einwenden, dass Annahmen des Messmodells (7.45) nicht erfüllt seien, insbesondere die Annahme gleicher Itemladungen. Wie in Abschnitt 5.2.2 ausgeführt, wird in der Generalisierbarkeitstheorie aufgrund der Annahme des repräsentativen Sampling postuliert, dass alle Items gleichgewichtet die Population repräsentieren. Eine latente Variable $\theta_p = \nu_p$ entsteht also erst durch diese Annahme. Dies steht im Gegensatz zu Faktorenanalysen und IRT-Modellen, in denen die Existenz der latenten Variablen für eine feste Itemmenge postuliert wird und Itemladungen als Itemparameter geschätzt werden. Demzufolge ist die Frage nach dem Modellfit in der Generalisierbarkeitstheorie nicht wohldefiniert.

Glas (2012b) spricht sich dafür aus, die Generalisierbarkeitstheorie in die Item-Response-Theorie zu integrieren. Das lineare Modell (7.46) der G-Theorie ist als Varianzkomponentenmodell (zumindest deskriptiv) auch für dichotome Daten anwendbar[12]. Glas (2012b) schreibt das Modell der G-Theorie mit Hilfe der logistischen Linkfunktion als

$$\text{logit}\, P(X_{pi} = 1) = \mu + \nu_p + \nu_i \tag{7.47}$$

In Analogie zur linearen Linkfunktion (7.46) können wiederum Varianzkomponenten σ_p^2 und σ_i^2 bestimmt werden, die nun allerdings nicht in der Metrik von Lösungshäufigkeiten (der identischen bzw. absoluten Metrik), sondern in der Logitmetrik interpretiert werden müssen. Wie in Abschnitt 7.1.1 mehrfach argumentiert, wird in Modellen mit der logistischen Linkfunktion die Fehlervarianz auf die Varianz der logistischen Verteilung fixiert. Dies hat als Konsequenz, dass bei Einfügen weiterer fester und zufälliger Effekte in (7.47) die Varianzkomponente σ_p^2 nicht mehr über verschieden komplexe Modelle hinweg

[12]Natürlich wäre eine Maximum Likelihood basierte Schätzung mit Normalverteilungsannahme für dichotome Items im Allgemeinen fehlspezifiziert. Dieser Befund wird aber fälschlicherweise häufig herangezogen, um die Inadäquatheit von linearen Modellen für dichotome Daten zu begründen. Ähnliche Diskussionen werden bei der Abgrenzung des linear probability model und der logistischen Regression für dichotome abhängige Variablen geführt (Angrist & Pischke, 2008).

vergleichbar ist. Damit ändert sich auch die Metrik und die Bedeutung von Personenparametern, die man aus diesen Modellen der G-Theorie mit der logistischen Linkfunktion extrahiert. Diese Überlegungen bedeuten nicht, dass Modelle (7.47) mit der logistischen Linkfunktion keinen Einsatz finden sollten. Umgekehrt argumentiere ich in Abschnitt 3.2, dass das Rasch-Modell als G-Theorie-Modell (bzw. Domain Sampling repräsentierend) aufgrund der Gleichgewichtung der Items aufgefasst werden kann (siehe auch Robitzsch et al., 2015). Eine mögliche Verletzung der lokalen stochastischen Unabhängigkeit wird dabei bewusst in Kauf genommen, weil diese Annahme nur die „Bestapproximation“ mit austauschbaren Items für das unabhängige Sampling von Items repräsentiert (vgl. auch Robitzsch & Lüdtke, 2015).

Verbindung von Generalisierbarkeitstheorie und Item-Response-Theorie

Briggs und Wilson (2007) versuchen die Ansätze der Item-Response-Theorie und der Generalisierbarkeitstheorie zu verbinden. Grob gesprochen wird dabei ein IRT-Modell mit einer nichtlinearen Linkfunktion angepasst und die Ergebnisse in die Metrik der linearen (d.h. identischen) Linkfunktion transformiert. Dabei wird zunächst das Rasch-Modell als statistisches Modell angenommen

$$P(X_{pi} = 1) = \Psi(\theta_p - b_i) = P(\theta_p, b_i) \tag{7.48}$$

wobei Ψ die logistische Linkfunktion bezeichnet und $\hat{X}_{pi} = P(\theta_p, b_i)$ die vorhergesagte Wahrscheinlichkeit bezeichnet. Das Modell (7.48) wird unter der Annahme zufälliger Personeneffekte θ_p und zufälliger Itemeffekte b_i (Itemschwierigkeiten) angepasst[13]. Eine der beiden Verteilungen wird dabei als zentriert angenommen. Dieses Modell kann beispielsweise in der Software WinBUGS mit MCMC-Methoden (Lunn, Thomas, Best & Spiegelhalter, 2000; Lunn, Spiegelhalter, Thomas & Best, 2009) angepasst werden, die für jede Iteration t in einer MCMC-Kette Realisierungen für θ_p und b_i liefert. Aus diesen Realisierungen kann dann die Vorhersage $\hat{X}_{pi} = P(\theta_p, b_i)$ direkt berechnet werden. Die Verbindung zwischen der Item-Response-Theorie und der Generalisierbarkeitstheorie ist dadurch hergestellt, dass sich die vorhergesagten Wahrscheinlichkeiten in der originalen Metrik der dichotomen Item Responses befinden, die in der G-Theorie verwendet wird. Der Gesamtmittelwert μ im linearen Modell der G-Theorie berechnet sich gemäß

$$\mu = \int_\theta \int_b P(\theta, b) f_\theta(\theta) f_b(b) \, \mathrm{d}\theta \, \mathrm{d}b \approx \frac{1}{N} \cdot \frac{1}{I} \sum_{p=1}^{N} \sum_{i=1}^{I} P(\theta_p, b_i) \tag{7.49}$$

Dabei bezeichnen f_θ und f_b die normalverteilten Dichtefunktion der zufälligen Effekte des Rasch-Modells (7.48). Das Integral in (7.49) deutet an, dass die erwarteten Wahrscheinlichkeiten „ausgemittelt“ werden. In einem MCMC-Verfahren entspricht das Ausmitteln einfach der Berechnung des Mittelwertes $P(\theta_b, b_i)$ über alle Personen $p = 1, \ldots, N$ und Items $i = 1, \ldots, I$. Der zentrierte Personeneffekt ν_p wird durch Integration (Mittelung)

[13] IRT-Modelle mit hierarchischen Verteilungen auf Itemparametern (Fox, 2010) werden von Brennan (2006) als „IRT-Version“ der G-Theorie angesehen.

über alle Items erhalten, d.h. man berechnet

$$\nu_p = \int_b P(\theta, b) f_b(b)\,\mathrm{d}b - \mu \approx \frac{1}{I}\sum_{i=1}^{I} P(\theta_p, b_i) - \mu \tag{7.50}$$

Für die Itemeffekte ν_i berechnet man in Analogie

$$\nu_i = \int_\theta P(\theta, b) f_\theta(\theta)\,\mathrm{d}\theta - \mu \approx \frac{1}{N}\sum_{p=1}^{N} P(\theta_p, b_i) - \mu \tag{7.51}$$

Die Personeneffekte θ_p und Itemeffekte b_i der Logitmetrik des Rasch-Modells werden also nichtlinear auf die identische Metrik der G-Theorie in die Effekte ν_p und ν_i transformiert. Zusätzlich kann man für die erwarteten Häufigkeiten $P(\theta_p, b_i)$ noch Interaktionen ν_{pi} definieren:

$$\nu_{pi} = P(\theta_p, b_i) - \mu - \nu_p - b_i \tag{7.52}$$

Diese Interaktion ν_{pi} kennzeichnet, wie gut sich die Wahrscheinlichkeiten $P(\theta_p, b_i)$ additiv zerlegen lassen. Es ist klar, dass Additivität in der Logitmetrik des Rasch-Modells gilt, jedoch damit nicht automatisch in der transformierten Metrik der G-Theorie. Durch Integration über θ bzw. b lassen sich Varianzkomponenten $Var(\nu_p)$, $Var(\nu_i)$ und $Var(\nu_{pi})$ gewinnen. Aufgrund der Zentriertheit von ν_p gilt beispielsweise

$$\sigma_p^2 = Var(\nu_p) = \int_\theta \nu_p^2 f_\theta(\theta)\,\mathrm{d}\theta \approx \frac{1}{N}\sum_{p=1}^{N} \nu_p^2 \tag{7.53}$$

Das transformierte Modell der G-Theorie lautet damit insgesamt

$$X_{pi} = \mu + \nu_p + \nu_i + \nu_{pi} + e_{pi} \tag{7.54}$$

Gegenüber dem Modell (7.45) tritt also noch der zusätzliche Term ν_{pi} auf. Passt man jedoch das lineare Modell (7.45) an, so sind die Varianzen von ν_{pi} und e_{pi} konfundiert. Ich argumentiere, dass es fragwürdig ist, weshalb zunächst ein Modell in der Logitmetrik angepasst wird, um im Anschluss besser interpretierbare Ergebnisse in der transformierten linearen Metrik zu erhalten. Eher sollte man vor dem Hintergrund einer konkreten Fragestellung entscheiden, ob man die Varianzkomponenten der G-Theorie in der linearen Metrik im Sinne von Brennan (2001a) oder der logistischen Metrik im Sinne von Glas (2012a) interpretieren will. Im Hinblick auf Varianzaufklärungsmaße ist dabei die lineare Metrik der G-Theorie einfacher, auch wenn die in der Literatur Varianzaufklärungsmaße für Mehrebenenmodelle mit der logistischen Linkfunktion vorgeschlagen werden (Goldstein, Browne & Rasbash, 2002). Die (am einfachsten zu interpretierenden) Vorschläge beruhen allerdings wiederum auf dem Ansatz, die aufgeklärte Varianz in der linearen Metrik (d.h. der Originalmetrik) zu berichten.

Verschiedenheit von Generalisierbarkeitstheorie (bzw. klassischer Testtheorie) und Item-Response-Theorie

Auch wenn die Arbeiten von Glas (2012a) und Briggs und Wilson (2007) strukturelle Ähnlichkeiten zwischen der G-Theorie und der IRT betonen, so hebt Brennan (2006) die

Unterschiede beider Ansätze heraus (siehe auch Blömeke, Gustafsson & Shavelson, 2015). Während die Item-Response-Theorie auf Items fokussiert, analysieren die klassische Testtheorie (KTT) und die G-Theorie primär die Eigenschaften des Tests (Brennan, 2006; siehe auch Abschnitt 5.2.2). Die IRT wird dabei von Brennan (2006) als *scaling model*, die G-Theorie als *measurement model* aufgefasst. Sowohl die latente Variable als auch der Fehler wird dabei in der G-Theorie unter der Annahme von austauschbaren und unendlich vielen Items durch die manifesten Items mit Erwartungswerten bzw. Varianzen definiert. Die IRT hingegen betrachtet Eigenschaften des Tests mit festen Items – nämlich die im Test eingesetzten Items – und generalisiert die Befunde nicht auf eine größere Itemmenge. Demzufolge wird die Existenz der latenten Variablen in der IRT postuliert. Brennan (2006) merkt an, dass es keinen Fehlerterm „per se“ in der IRT gäbe. Dies könnte man dadurch begründen, dass die Fehlervarianz der latenten Item Responses (und damit die Linkfunktion) in der IRT fixiert wird (siehe Abschnitt 7.1.1). Bedingte Standardmessfehler in der IRT sind daher auf die postulierte latente Variable θ bedingt. Modellfehler (Modell-Misfit) gehen dabei nicht in die statistische Inferenz für Personenfähigkeiten in der IRT ein. In der G-Theorie werden diese jedoch in der Fehlervariablen e_{pi} abgebildet (Brennan, 2006; siehe auch Abschnitt 5.2.2).

Zusammenfassend merkt Brennan (2004, 2006, 2011) an, dass die IRT gegenüber der G-Theorie (und der KTT) stärkere Annahmen trifft, da Items in der IRT nicht als austauschbar angenommen werden.

Die Modellgleichungen der KTT der Form $X_i = T_i + E_i$ für Items i sind tautologisch und lassen sich zunächst ohne Annahmen an die Dimensionaliät der true scores $T_1, \ldots, T_I$ formulieren (Sijtsma & van der Ark, 2015). Davon wird die auf einem Faktormodell (konfirmatorische ein- oder mehrdimensionale Faktorenanalyse) basierende Bestimmung einer Reliabilität von der klassischen Reliabilität unterschieden (siehe ebd.). Ein Teil der psychometrischen Literatur sieht jedoch die KTT als spezielle Faktormodelle wie die Modelle tau-äquivalenter Messungen oder tau-kongenerischer Messungen an (Steyer & Eid, 2001; Lewis, 2007; Raykov & Marcoulides, 2010). Unter dieser Perspektive erscheinen IRT und KTT praktisch äquivalent und unterscheiden sich nur in den Annahmen über die Linkfunktionen (und Verteilungen der Residuen). Ich schließe mich allerdings eher der Perspektive von Brennan an und betonen die konzeptuellen Unterschiede von IRT und KTT (bzw. G-Theorie).

7.3.3 Domain Sampling Interpretation von Cronbachs Alpha

In diesem Abschnitt diskutieren wir verschiedene Interpretationen der Definition einer Reliabilität. Dabei kann man grob zwischen designbasierten und modellbasierte Ansätze der Reliabilitätsbestimmung unterscheiden, auch wenn sich rein formal designbasierte Ansätze in modellbasierte Ansätze überführen lassen (siehe Abschnitt 5.2).

In der Literatur findet man häufig die Aussage, dass für die Verwendung von Cronbachs Alpha (Cronbach, 1951) als Reliabilitätsmaß die Annahme eines tau-äquivalenten Messmodells gegeben sein muss. Das heißt, dass die Skala eindimensional ist, alle Itemladungen identisch sind und die residualen Fehlervariablen unkorreliert sind (siehe z.B. Cho & Kim, 2015; Gignac, 2014; Raykov & Marcoulides, 2010; Steyer & Eid, 2001). Dies ist aber nur korrekt, wenn man die Reliabilitätsdefinition nur modellbasiert (d.h. auf Basis

eines passenden Messmodells) vornehmen möchte.

Formal ist Cronbachs Alpha α (Cronbach, 1951; siehe auch Cronbach & Shavelson, 2004) für beobachtete Kovarianzen s_{ij} von Itempaaren (i, j) definiert durch

$$\alpha = \frac{I\overline{c}_I}{\overline{v}_I + (I-1)\overline{c}_I} \tag{7.55}$$

wobei $\overline{v}_I$ die mittlere Varianz aller I Items und $\overline{c}_I$ die mittlere Kovarianz aller Itempaare bezeichnet. Input für die Berechnung von α sind also nur Kovarianzen, wobei diese nur als suffiziente Statistiken $\overline{c}_I$ und $\overline{v}_I$ eingehen. Cronbach (1951) zeigt, dass α gleich dem Mittelwert aller möglichen Testhalbierungs-Reliabilitäten (*split half* Reliabilitäten) ist.

Domain Sampling Interpretation von Alpha

Die Definition (7.55) kann jedoch auch unter einer (designbasierten) Domain Sampling Perspektive verstanden werden (Tryon, 1957; siehe auch Hulin et al., 2001). Sei dazu ein Test mit einer Kovarianzmatrix $\boldsymbol{S}$ gegeben und einem Summenwert $S = X_1 + \ldots + X_I$ gegeben. Mit der Kenntnis von $\boldsymbol{S}$ sind auch die Cronbachs Alpha definierenden Größen $\overline{c}_I = f_c(\boldsymbol{S})$ und $\overline{v}_I = f_v(\boldsymbol{S})$ bekannt. Tryon (1957) argumentiert, dass man die Reliabilität α die Korrelation des Tests mit zugehöriger Kovarianzmatrix $\boldsymbol{S}$ mit einem gleich langen „parallelen“ Test mit ähnlichen Eigenschaften und dzugehöriger Kovarianzmatrix $\tilde{\boldsymbol{S}}$ ist. Dabei wird Ähnlichkeit dadurch definiert, dass $\tilde{\boldsymbol{S}}$ dieselbe mittlere Kovarianzmatrix $\overline{c}_I$ und dieselbe mittlere Varianz $\overline{v}_I$ wie $\boldsymbol{S}$ besitzt. Klarerweise stimmt damit die Reliabilität α für beide Tests überein. Es wird demzufolge nach Tryon (1957) nicht postuliert, dass alle Items dieselben Ladungen besitzen, wie dies in der o.g. Literatur der modellbasierten Begründung für Alpha gefordert wird. Vielmehr wird eine Repräsentativitätsannahme postuliert, dass sich eine hypothetische Testreplikation auf Items mit ähnlichen Eigenschaften bezieht (Brennan, 2001c). Sowohl die Items des originalen Tests als auch die Items in der Testreplikation sind daher geeignete „Zufallsziehungen“ aus der (ggf. hypothetischen) Population aller Items (d.h. der Itemdomäne).

In Abschnitt 7.3.1 wurde die Idee des Domain Samplings im Hinblick auf die statistische Inferenz diskutiert. Dabei wird die Verteilung der Realisierungen in der $N \times I$-Datenmatrix $\boldsymbol{X}$ mit Item Responses im Hinblick auf eine unendliche Personenpopulation ($N \to \infty$) und eine unendliche Itempopulation ($I \to \infty$) diskutiert (siehe Cronbach & Shavelson, 2004). Für eine konkreten Test mit I Items ist dabei bei fest gehaltenen Items die empirische Kovarianzmatrix $\boldsymbol{S} = \boldsymbol{S}_I$ eine Schätzung der Kovarianzmatrix $\boldsymbol{\Sigma}_I$ in der Population aller Personen. Diese Matrix für I Items entsteht aber durch Sampling aus einer unendlichen Itempopulation, die man mit einer (abzählbar) unendlichdimensionalen Matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_\infty$ assoziieren kann[14]. Für diese Matrix $\boldsymbol{\Sigma}$ sollen die mittlere Kovarianz γ_c und die mittlere Varianz γ_v existieren. Für die Population aller Items kann man daher für einen hypothetischen Test mit I Items die Populationskenngröße von Cronbachs Alpha definieren

$$\alpha_{pop} = \frac{I\gamma_c}{\gamma_v + (I-1)\gamma_c} \tag{7.56}$$

[14]Dabei lässt man in $\boldsymbol{\Sigma}_I$ für beide Dimensionen I gegen Unendlich gehen.

Eine übliche statistische Inferenz im Hinblick auf Item Sampling (gleich Domain Sampling) beruht dann auf der Idee, dass $\overline{c}_I$ eine Schätzung für γ_c und $\overline{v}_I$ eine Schätzung für γ_v ist und daher α eine Schätzung für das in der Itempopulation definierte Cronbachs Alpha α_{pop} darstellt[15]. Im Hinblick auf Item Sampling werden daher die Größen $\overline{c}_I$ und $\overline{v}_I$ Standardfehler besitzen.

Einige Autoren definieren einen *Domain Score* $D_p = \lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{M} X_{pi}$ als wahren Wert eines Tests (Nunnally & Bernstein, 1994; Revelle, 2015, Kap. 7). Cronbachs Alpha wird dann als quadrierte Korrelation eines Tests mit I Items aus dieser Itemdomäne und dem Domain Score abgeleitet. Wir merken an, dass diese Ableitung nur die Annahme trifft, dass alle Itemvarianzen endlich sind und keinerlei Annahmen an die Faktorenstruktur des Tests getroffen werden muss.

Wird die Itemdomäne durch ein K-dimensionales Faktormodell beschrieben, so zeigt McDonald (1978), dass die Korrelation eines Testscores mit dem Domain Score durch McDonald's Omega ω_t (siehe auch McDonald, 1999; Zinbarg et al., 2005) gegeben ist. Im Allgemeinen gilt $\alpha \leq \omega_t$. Dies begründet die Empfehlungen einiger Autoren das Maß ω dem Maß α vorzuziehen (siehe z.B. Peters, 2014 oder Cho & Kim, 2015), da α nur eine untere Schranke der Reliabilität ist (vgl. auch Revelle & Zinbarg, 2009). Die Annahmen für die Bestimmung von α und ω unterscheiden sich jedoch: Im ersten Fall von α erfolgt eine Ableitung auf Basis einer designbasierten Reliabilität, im zweiten Fall von ω erfolgt die Bestimmung der Reliabilität modellbasiert mit Hilfe eines mehrdimensionalen Faktormodells.

Ein Varianzkomponentenmodell für Kovarianzen zur Bestimmung des Standardfehlers von Alpha unter Item Sampling

Im Folgenden beschreiben wir ein einfaches statistisches Modell für die statistische Inferenz von Cronbachs Alpha unter Domain Sampling. Für den Test mit I Items folgt dabei die empirische Kovarianz $\boldsymbol{S}_I$ einer Invers-Wishart-Verteilung mit der Populationsmatrix $\boldsymbol{\Sigma}_I$ und N Freiheitsgraden (Anderson, 2003). Durch Sampling von Personen entsteht also ein Fehler e_{ij} für Itempaare (i, j), so dass sich die empirische Kovarianz schreiben lässt als

$$s_{ij} = \sigma_{ij} + e_{ij} \tag{7.57}$$

Der Mittelwert aller Kovarianzen σ_{ij} (für $i \neq j$) in der Population aller Items ist γ_c. Wir nehmen als datengenerierendes Modell für die unendliche Itempopulation an, dass

$$\sigma_{ij} = \gamma_c + u_i + u_j + \epsilon_{ij} \tag{7.58}$$

mit zentrierten normalverteilten unabhängigen Itemeffekten u_i und Residuen ϵ_{ij}. Die unendlichdimensionale Kovarianzmatrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_\infty$ wird dabei in Verteilung durch wenige Parameter in (7.58) charakterisiert. Dabei werden γ_c sowie $Var(u_i) = \sigma_u^2$ und $Var(e_{ij}) = \sigma_\epsilon^2$ geschätzt. Für die Schätzung der mittleren Itemkovarianz $\overline{c}_I$ ergibt sich dann der durch Item Sampling bedingte Standardfehler

$$SD(\overline{c}_I) = \frac{1}{2} \cdot \frac{\sigma_u}{\sqrt{I}} + \frac{\sigma_\epsilon}{\sqrt{I(I-1)/2}} \tag{7.59}$$

[15] Eine entsprechende Aussage der Konvergenz lässt sich mit dem starken Gesetz der Großen Zahlen (Wasserman, 2004) begründen.

Für die Bestimmung der Präzision der mittleren Itemkovarianz hat Cortina (1993) einen ähnlichen Ansatz verfolgt, ignoriert allerdings den Varianzterm σ_u^2 und geht daher von unabhängigen Itempaaren aus. Dies ist unseres Erachtens keine plausible Annahme[16].

Das Varianzkomponentenmodell (7.58) lässt sich mit einer Quadratsummen-Methode, Maximum-Likelihood-Verfahren oder MCMC-Verfahren einfach schätzen. Wir merken an, dass das Modell (7.58) starke Ähnlichkeiten mit dem *social relations model* (Warner, Kenny & Stoto, 1979; siehe auch Lüdtke, Robitzsch, Kenny & Trautwein, 2013) besitzt, wobei allerdings weniger Parameter zu schätzen sind.

Alpha unter der Perspektive der Generalisierbarkeitstheorie

Eine statistische Inferenz für die abgeleitete Größe Cronbachs Alpha $\alpha = f(\overline{c}_I, \overline{v}_I)$ wäre dann mit der Delta-Formel (Wasserman, 2004) realisierbar. Unsere Überlegungen für $\overline{c}_I$ können dabei auf $\overline{v}_I$ übertragen werden. Alternativ kann die simultane statistische Inferenz für Personen und Items jedoch auch mit der Methode des Double Jackknife (Brennan, 2001a, Kap. 6) durchgeführt werden.

Cronbachs Alpha stimmt jedoch auch mit einem Reliabilitätsmaß der G-Theorie überein (Sijtsma & van der Ark, 2015). Es möge dabei wiederum die Grundgleichung der G-Theorie für Personen und Items gelten

$$X_{pi} = \mu + \nu_p + \nu_i + e_{pi} \tag{7.45}$$

Alpha ist dann gerade der Generalisierbarkeitskoeffizient für einen Test mit I Items, wenn die Varianz $Var(\nu_i)$ in der Berechnung ignoriert wird, da verschiedene Itemeffekte ν_i die Reihenfolge von Personen nicht ändern (siehe Sijtsma & van der Ark, 2015). Das Modell der G-Theorie sieht die Items als austauschbar an und modelliert daher wegen $Cov(e_{pi}, e_{pj}) = 0$ für verschiedene Items i und j homogene Varianzen, d.h.

$$Cov(X_i, X_j) = \sigma_{ij} = \sigma_p^2 \tag{7.61}$$

Gegenüber unserem vorgeschlagenen Varianzkomponentenmodell (7.58) für die Populationskovarianzmatrix modelliert man also in der G-Theorie das Sampling von Personen und Items simultan und behandelt Personen und Items als austauschbar. Unsere obige Ableitung für die Inferenz von α nimmt zunächst jedoch an, dass die Beobachtungen $\boldsymbol{X}$ aus einer Kovarianzmatrix $\boldsymbol{S}$ gesampelt werden und alle Einträge dieser Matrix in einem zweiten Schritt mit einem Varianzkomponentenmodell stochastisch modelliert werden. Das vorgeschlagene Modell beschreibt also die Kovarianzstruktur zwischen Items durch ein komplexeres statistisches Modell als in der G-Theorie.

[16] Nimmt man etwa ein *approximate factor model* (Chamberlain & Rothschild, 1983; siehe den nächsten Abschnitt 7.3.4) mit einer Dimension an, dann ist

$$\sigma_{ij} = \lambda_i \lambda_j + \delta_{ij} \tag{7.60}$$

wobei wir δ_{ij} als einen Modellfehler im tau-kongenerischen Modell interpretieren (siehe den nächsten Abschnitt). Schreibt man $\lambda_i = \sqrt{\gamma_c} + \nu_i$ und setzt in (7.60) ein, so erhält man eine additive Zerlegung der Form (7.58). Die Itemeffekte u_i sind auch dann bedeutsam, wenn das eindimensionale Faktormodell exakt gilt, d.h. $\sigma_{ij} = \lambda_i \lambda_j$. In diesem Fall ergibt sich ein Modellansatz $\sigma_{ij} = \gamma_c + u_i + u_j$, in dem der Interaktionsterm ϵ_{ij} in (7.58) gleich Null ist.

Legitimation von Alpha ohne Domain Sampling Annahme

Lässt sich Cronbachs Alpha als Reliabilitätsmaß auch legitimieren, wenn ein Test mit I festen Items vorliegt und keine Generalisierung auf eine größere Itempopulation angestrebt wird? In Abschnitt 5.2.1 sind wir von der Gleichung $\boldsymbol{X} = \boldsymbol{T} + \boldsymbol{E}$ der klassischen Testtheorie (KTT) in der multivariaten Version für I Items ausgegangen. Die Kovarianzmatrix $\boldsymbol{\Sigma}$ von $\boldsymbol{X}$ lässt sich zerlegen in eine Kovarianzmatrix $\boldsymbol{\Phi}$ der wahren Werte und eine Kovarianzmatrix $\boldsymbol{\Theta}$, so dass $\boldsymbol{\Sigma} = \boldsymbol{\Phi} + \boldsymbol{\Theta}$. Wie die Grundgleichung der KTT ist auch diese Matrixzerlegung tautologischer Natur, da ohne strukturelle Annahmen keine eindeutige Zerlegung existiert. Die (klassische) Reliabilität ρ nach Lucke (2005) wurde in Abschnitt 5.2.1 mit Hilfe der Matrixzerlegung $\boldsymbol{\Sigma} = \boldsymbol{\Phi} + \boldsymbol{\Theta}$ definiert als

$$\rho = \frac{\mathbf{1}'\boldsymbol{\Phi}\mathbf{1}}{\mathbf{1}'\boldsymbol{\Phi}\mathbf{1} + \mathbf{1}'\boldsymbol{\Theta}\mathbf{1}} \tag{5.4}$$

In Abschnitt 5.2.2 haben wir dann angenommen, dass $\boldsymbol{\Phi} = \phi\mathbf{1}\mathbf{1}'$ gilt. Die Kovarianzmatrix der Fehler $\boldsymbol{E}$ kann dann durch $\boldsymbol{\Theta} = \boldsymbol{\Sigma} - \boldsymbol{\Phi}$ bestimmt werden. Die spezielle Annahme $\boldsymbol{\Phi} = \phi\mathbf{1}\mathbf{1}'$ bedeutet, dass im Hinblick auf die wahren Werte alle Items als austauschbar betrachtet werden *sollen*, da dieselbe Kovarianz zwischen den wahren Werten angenommen wird. Die Modellspezifikation entspricht dabei dem Modell tau-äquivalenter Messungen mit korrelierten Residuen, wobei eine Identifikationsannahme an die Schätzung gestellt werden muss (hier soll die Summe der Nichtdiagonalelemente in $\boldsymbol{\Theta}$ gleich Null sein). Dann entspricht die klassische Reliabilität ρ genau Cronbachs Alpha α (siehe Abschnitt 5.2.2). Man wird einwenden können, dass die Annahme gleicher Itemladungen unplausibel sei und praktisch nicht erfüllt sein wird. Bei der Spezifikation der Matrixzerlegung $\boldsymbol{\Sigma} = \boldsymbol{\Phi} + \boldsymbol{\Theta}$ ist dieser Einwand aber irrelevant, denn die verschiedenen Itemladungen werden als Fehler in $\boldsymbol{\Theta}$ abgebildet. Man nimmt aufgrund der Bedingung $\boldsymbol{\Phi} = \phi\mathbf{1}\mathbf{1}'$ an, dass alle I Items im Hinblick auf den wahren Wert der Skala gleichgewichtet werden *sollen*, jede Abweichung von einer Gleichgewichtung führt dazu, dass die feste Itemmenge nicht mehr „repräsentativ“ für das Konstrukt ist. Für die Erfassung von Mathematikkompetenzen argumentieren Robitzsch et al. (2015) im Kontext der Gleichgewichtung im Rasch-Modell, dass abweichende Itemgewichtungen von der Gleichgewichtung zu einer Regewichtung der Bedeutung einzelner mathematischer Kompetenzbereiche im Hinblick auf die Gesamtskala führen und daher theoretischen Annahmen der Bedeutung von Itemgruppen widersprechen.

7.3.4 Mehrdimensionale Faktormodelle

Im folgenden Abschnitt wird beleuchtet, welche Bedeutung das Domain Sampling für Faktormodelle besitzt. Da man IRT-Modelle als Faktormodelle mit einer logistischen Linkfunktion auffassen kann (Takane & De Leeuw, 1987), gelten unsere Überlegungen gleichermaßen auch für IRT-Modelle. Wir grenzen dabei den Begriff des Modellfehlers vom Domain Sampling ab. Zunächst führen wir dabei den Gedanken des Modellfehlers am Beispiel der linearen Regression ein. Im zweiten Teil stellen wir das approximative Faktormodell (approximate factor model) dar. Im dritten Teil kontrastieren wir diesen Ansatz mit einem aktuell vorgeschlagenen Ansatz der stochastischen Modellierung von Modellfehlern nach Wu und Browne (2015). Im vierten Teil konkretisieren wir unsere

Überlegungen in Anwendung auf IRT-Modelle. Im letzten Teil diskutieren wir Bayesianische Ansätze von Faktorstrukturen, die mittels Priorverteilungen Modellfehler in die Analyse einbeziehen.

Modellfehler in der linearen Regression

Für die Illustration der Bedeutung von Modellfehlern betrachten wir den Fall der linearen Einfachregression. Wir halten uns dabei an die Darstellung in Berk et al. (2014). Dabei liegen Beobachtungen (Y_p, X_p) der abhängigen Variablen Y und der unabhängigen Variablen X vor. Die Kovariate X möge dabei einem zufälligen Design folgen (vgl. Berk et al., 2014). Die bedingte Erwartung $\mu(X) = E(Y|X)$ wird dabei als Regression in der Population definiert. Diese Regressionsfunktion wird dabei im Allgemeinen nichtlinear sein. In Anwendungen beschreibt man den Zusammenhang zwischen X und Y häufig durch einfache funktionale Formen. Im Fall einer linearen Regression nimmt man $\omega(X) = \beta_0 + \beta_1 X$ an, so dass die Regressionskoeffizienten β_0 und β_1 zu schätzen sind.

Für die konkrete Stichprobe nimmt man an, dass die Daten der Modellgleichung

$$Y_p = \mu(X_p) + \varepsilon_p = \omega(X_p) + \delta(X_p) + \varepsilon_p \quad \text{mit} \quad E(\varepsilon_p) = 0 \tag{7.62}$$

folgen, wobei der Modellfehler $\delta(X)$ durch $\delta(X) = \mu(X) - \omega(X)$ gegeben ist. Wir merken an, dass der Modellfehler im Allgemeinen auch für einen großen Stichprobenumfang (von Personen) nicht verschwindet, da dieser Fehler den Fehler der linearen Approximation $\omega(X)$ für die nichtlineare Funktion $\mu(X)$ quantifiziert.

Berk et al. (2014) argumentieren, dass der Einsatz der linearen Regression auch für nichtlineare Funktionsverläufe legitimiert ist. Die Regressionskoeffizienten beschreiben dann „mittlere“ lineare Einflüsse der Prädiktoren $\boldsymbol{X}$. Diese mittleren Effekte können auch dann bedeutsam sein, wenn der erwartete Einfluss von X auf Y nichtlinear ist. In der kausalen Inferenz interessiert man sich auch dann häufig für *average treatment effects* (ATE) einer Treatmentvariablen T, wenn diese Effekte in Abhängigkeit von Kovariaten variieren (Morgan & Winship, 2007). Der ATE beschreibt dann den mittleren Effekt in der gesamten Population.

Für das Regressionsmodell (7.62) können klassische Maße der Varianzaufklärung bestimmt werden. Der Anteil R^2 der aufgeklärten Varianz in der Stichprobe wird mittels

$$R^2 = \frac{Var(\omega(X))}{Var(Y)} = 1 - \frac{Var(\delta(X)) + Var(\varepsilon)}{Var(\omega(X)) + Var(\delta(X)) + Var(\varepsilon)} \tag{7.63}$$

bestimmt. Sowohl Stichprobenfehler (unsystematische Varianz $Var(\varepsilon)$ als auch Modellfehler $Var(\delta(X))$ werden dann als Fehlervarianz angesehen. Man kann allerdings auch die Bedeutung des Modellfehlers $\delta(X)$ als Approximationsfehler der bedingten Erwartung $\mu(X) = E(Y|X) = \omega(X) + \delta(X)$ durch die lineare Regression $\omega(X)$ definieren. Die Varianzaufklärung in der Population R^2_{pop} ist dann gegeben durch

$$R^2_{pop} = \frac{Var(\omega(X))}{Var(\mu(X))} = 1 - \frac{Var(\delta(X))}{Var(\omega(X)) + Var(\delta(X))} \tag{7.64}$$

Ist der Modellfehler (d.h. der Approximationsfehler) der linearen Regression gleich Null (d.h. $\delta(X) \equiv 0$), dann ist $R^2_{pop} = 1$. Das lineare Regressionsmodell ist demnach korrekt spezifiziert, wenn R^2_{pop} den Wert Eins annimmt. Für sehr viele Beobachtungen (X_p, Y_p) lässt

sich $\mu(X)$ nichtparametrisch schätzen und der „mittlere Approximationsfehler“ $Var(\delta(X))$ ist bestimmbar (Fox, 1997)[17].

Die klassische statistische Inferenz (d.h. für Standardfehler) ist jedoch bei der Existenz von Modellfehlern nicht mehr korrekt (Berk et al., 2014). Robuste Standardfehler (Angrist & Pischke, 2008) oder Resampling-Verfahren wie Jackknife oder Bootstrap führen jedoch auch bei Existenz von Modellfehlern zu einer validen statistischen Inferenz (Berk et al., 2014).

In der Literatur zur *uncertainty quantification* werden häufig Regressionsmodelle mit Modellfehlern vom Typ (7.62) diskutiert (siehe Kennedy & O'Hagan, 2001; Brynjarsdóttir & O'Hagan, 2014; vgl. auch MacCallum & O'Hagan, 2015). Für den Modellfehler $\delta(X)$ wird dabei ein Gaußscher Prozess mit einem Erwartungswert von Null und korrelierten Fehlern x_k bzw. x_m angenommen, wobei die Korrelation eine Funktion der Differenz $|x_k - x_m|$ sein kann. In diesem Sinn beschreibt der Modellfehler $\delta(X)$ systematische Variation, während die Varianz der Residuen $Var(\varepsilon_p)$ unsystematische Fehler beschreibt.

Die Bedeutung von möglichen „Modellverletzungen“ in der Regression hängt davon ab, ob man die Prädiktoren als fest oder als zufällig interpretiert. Betrachtet man die Kovariate X als fest, so kann man für feste Ausprägungen x_1 und x_2 prüfen, ob die Regressionsresiduen $Y - E(Y|X = x_1)$ und $Y - E(Y|X = x_2)$ unkorreliert sind, so dass man diese Modellannahme direkt prüfen (oder testen) kann. Im Fall zufälliger Kovariaten nimmt man nur an, dass die Residuen im Mittel einen Erwartungswert von Null haben, was per Konstruktion gegeben ist. Es ist dann nicht unterscheidbar, ob nichtlineare Abweichungen als ein nichtmodellierter bedingten Effekt oder als ein korrelierter Fehler interpretiert werden. Im ersten Fall der nichtmodellierten Nichtlinearität liegt eine Fehlspezifikation der bedingten Erwartung vor, so dass das Regressionsmodell nicht *first order correct* ist (Berk et al., 2014, S. 426). Im zweiten Fall der korrelierten Residuen ist die Annahme unkorrelierter Residuen verletzt, so dass das Modell nicht *second order correct* ist (Berk et al., 2014, S. 426).

Anstelle der Maximum Likelihood Schätzung oder der Kleinsten-Quadrate-Schätzung des Regressionsmodells (7.62) kann man auch robuste Schätzungen einsetzen (Maronna et al., 2006). Anstelle der Summe der Quadrate der Residuen $\sum_p e_p^2$ kann dabei beispielsweise auch die Summe der Beträge $\sum_p |e_p|$ oder andere robuste Zielfunktionen ρ mit $\sum_p \rho(e_p)$ betrachtet werden (siehe Fox, 1997, Kap. 14). Die Summe der Quadrate entspricht dabei der Normalverteilungsannahme der Residuen, die Summe der Beträge oder anderer robuster Zielfunktionen liefert effizientere Schätzer bei Abweichungen von der Normalverteilung, d.h. bei der Existenz von Ausreißern oder stark in den Rändern (*tails*) besetzten Verteilungen (z.B. der Laplaceverteilung, siehe Song, Yao & Xing, 2014).

Wir werden im Folgenden die Ideen des Modellfehlers der linearen Regression auf die Anpassung von Faktormodellen übertragen.

Faktormodelle und Modellfehler

Für einen Test mit I kontinuierlichen (zentrierten) Items $\boldsymbol{X}$ möge nun eine empirische Kovarianzmatrix $\boldsymbol{\Sigma}$ vorliegen. Für dichotome oder polytome Items sehen wir stattdessen

[17] Vgl. auch Douglas & Cohen, 2001 zur Beurteilung des Fits parametrisch spezifizierter Item-Response-Funktionen mit Hilfe nichtparametrisch geschätzter Item-Response-Funktionen.

$\boldsymbol{S}$ als Matrix tetrachorischer oder polychorischer Korrelationen an. In mehrdimensionalen Faktormodellen repräsentiert man die I Items durch K Faktoren $\boldsymbol{F}$ (latente Variablen). Residuale Fehlervariablen $\boldsymbol{E}$ werden meist als unkorreliert angenommen (Mulaik, 2009a).

Das lineare Faktormodell erfüllt dabei die Gleichung

$$\boldsymbol{X} = \boldsymbol{\Lambda F} + \boldsymbol{E} \quad \text{mit } Var(\boldsymbol{F}) = \boldsymbol{\Phi} \text{ und } Var(\boldsymbol{E}) = \boldsymbol{\Theta} \tag{7.65}$$

Dabei können in einer explorativen Faktorenanalyse (nahezu) alle Faktorladungen in der Matrix $\boldsymbol{\Lambda}$ frei geschätzt werden, in einer konfirmatorischen Faktorenanalyse gibt der Anwender vor, welche der Einträge geschätzt werden sollen. In den meisten Anwendungen ist die residuale Kovarianzmatrix $\boldsymbol{\Theta}$ eine Diagonalmatrix, was der Annahme der lokalen stochastischen Unabhängigkeit entspricht.

Die Populationskovarianzmatrix $\boldsymbol{\Sigma}$ schreibt sich dann mit (7.65) gemäß

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda \Phi \Lambda}^T + \boldsymbol{\Theta} \tag{7.66}$$

In die Modellanpassung der Faktorenanalysen gehen (bei vollständigen Daten) nicht die Rohdaten $\boldsymbol{X}$, sondern nur die suffiziente Statistik $\boldsymbol{S}$ ein. Das Sampling von Personen aus einer Population führt dazu, dass $\boldsymbol{S}$ eine Schätzung von $\boldsymbol{\Sigma}$ ist. In einem zweiten Schritt wird die Populationsmatrix durch einen niedrigdimensional(er)en Parameter gemäß des rechten Terms in (7.66) repräsentiert.

Die Modellgleichung (7.66) kann man jedoch auch einzeln für Items i und j notieren. Dann ist

$$\sigma_{ij} = \omega(\xi)_{ij} = \omega_{ij} = \sum_{k,k'} \lambda_{ik} \phi_{k,k'} \lambda_{jk'} + \theta_{ij} \tag{7.67}$$

Dabei haben wir die Bezeichnung $\omega(\boldsymbol{\xi})$ als Abkürzung eingeführt, wobei der Vektor $\boldsymbol{\xi}$ alle zu schätzenden Parameter beinhaltet. In vielen Anwendungen wird man aus Identifikationsgründen $\theta_{ij} = 0$ setzen, also von unkorrelierten Residuen ausgehen. Damit wird aber typischerweise ein (ggf. praktisch nicht relevanter) Modellfehler induziert (Cudeck & Henly, 1991; MacCallum & Tucker, 1991; MacCallum, 2003). Es existiert eine Modelldiskrepanz $\delta_{ij} = \sigma_{ij} - \omega_{ij}$ zwischen der Populationskovarianz σ_{ij} und der vorhergesagten Kovarianz ω_{ij} mit Hilfe des statistischen Modells. Wie im Fall der linearen Regression des vorhergehenden Abschnittes existieren demzufolge Modellfehler, die sich auch bei unendlich großem Stichprobenumfang von Personen nicht reduzieren lassen. Die beobachtete Kovarianz s_{ij} lässt sich dann schreiben als $s_{ij} = \sigma_{ij} + e_{ij}$ mit einem unsystematischen Samplingfehler e_{ij}. Der Modellfehler δ_{ij} ist von systematischer Natur. Insgesamt ergibt sich damit das Modell für die empirischen Kovarianzen

$$s_{ij} = \sigma_{ij} + e_{ij} = \omega_{ij} + \delta_{ij} + e_{ij} \tag{7.68}$$

Die Modellvorhersagen ω_{ij} werden aus dem zu schätzenden Parameter $\boldsymbol{\xi}$ berechnet.

Savalei (2014) betont, dass man sich die Anpassung von Faktormodellen als nichtlineare Regression von Kovarianzen s_{ij} auf modellimplizierte Kovarianzen vorstellen kann. In der linearen Regressionen sind die einzelnen Beobachtungen den Personen p zugeordnete Daten (X_p, Y_p). Im Faktormodell sind die „Beobachtungen“ durch die Itempaare (i, j) charakterisiert und die Daten sind durch die empirische Kovarianzmatrix $\boldsymbol{S} = (s_{ij})$ gegeben.

Die abhängige Variable Y_p in der Regression entspricht nun den beobachteten Kovarianzen s_{ij} im Faktormodell. Der unsystematische Fehler ε_p in der linearen Regression ist der durch Person Sampling verursachte Fehler e_{ij} in (7.68). Während in der linearen Regression die bedingte Erwartung $\mu(X) = E(Y|X = x)$ durch die lineare Regressionsfunktion $\omega(X)$ approximiert wird, betrachtet man im Faktormodell die Approximation der Kovarianz σ_{ij} in der Population mit Hilfe der nichtlinearen Regression $\omega_{ij} = \omega((i, j))$.

In der linearen Regression hatten wir außerdem das Varianzaufklärungsmaß R^2 in der Stichprobe eingeführt. Bezeichnen wir mit Var_{ij} die Variabilität über die Itempaare (i, j) (d.h. die (empirische) Varianz), so kann man für als Regression formulierte Faktormodelle ein analoges R^2 vorstellen (vgl. (7.63))

$$R^2 = \frac{Var_{ij}(\omega_{ij})}{Var_{ij}(s_{ij})} = 1 - \frac{Var_{ij}(\delta_{ij}) + Var_{ij}(e_{ij})}{Var_{ij}(\omega_{ij}) + Var_{ij}(\delta_{ij}) + Var_{ij}(e_{ij})} \tag{7.69}$$

Ggf. sollten die Ausgangsvariablen $\boldsymbol{X}$ zuvor standardisiert werden, damit die Modellabweichungen in einer ähnlichen Metrik interpretiert werden können. Die Varianzaufklärung R^2 gibt an, wie gut die empirische Kovarianzmatrix $\boldsymbol{S}$ (mit Einträgen s_{ij}) durch die modellimplizierte Kovarianzmatrix $\boldsymbol{\Omega}$ (mit Einträgen ω_{ij}) angepasst wird. Häufig interessiert man sich aber eher für die Frage, ob die Populationskovarianzmatrix $\boldsymbol{\Sigma}$ gut durch $\boldsymbol{\Omega}$ approximiert wird. Dies führt in Analogie zur Regression zur Varianzaufklärung R^2_{pop} in der Population (vgl. (7.64))

$$R^2_{pop} = \frac{Var_{ij}(\omega_{ij})}{Var_{ij}(\sigma_{ij})} = 1 - \frac{Var_{ij}(\delta_{ij})}{Var_{ij}(\omega_{ij}) + Var_{ij}(\delta_{ij})} \tag{7.70}$$

Dieses Maß quantifiziert die Bedeutung des Modellfehlers $\boldsymbol{\Delta} = (\delta_{ij})$. Das Faktormodell passt perfekt, falls $R^2_{pop} = 1$ erfüllt ist. In Faktormodellen wird daher im traditionellen Vorgehen ein nichtlineares Regressionsmodell spezifiziert, dass man auf $R^2_{pop} = 1$ testet. In konfirmatorischen Faktorenanalysen basieren Statistiken für den Modellmisfit (z.B. $RMSEA$ oder $SRMR$) häufig auf der Auswertung der Modellfehler δ_{ij} (Mulaik, 2009b). Praktisch ist dies äquivalent dazu, dass man nur eine kleine Varianz $Var_{ij}(\delta_{ij})$ der Modellfehler zulassen möchte. Die Statistik $SRMR$ könnte dabei auf Populationsebene gemäß $\sqrt{Var_{ij}(\delta_{ij})}$ definiert werden.

Verschiedene Schätzmethoden zur Anpassung von Faktormodellen wie in (7.68) diskutieren MacCallum et al. (2007). Dabei zeigen die Autoren, dass die verschiedenen Schätzmethoden zu unterschiedlichen Ergebnissen in Abhängigkeit von Annahmen über Modellfehler δ_{ij} und Stichprobenfehler e_{ij} gelangen. Die *unweighted least squares* (ULS) Schätzmethode bestimmt Modellparameter $\boldsymbol{\xi}$ als Minimum von

$$F_{ULS}(\boldsymbol{\xi}) = \sum_{i,j} (s_{ij} - \omega_{ij})^2 = \sum_{i,j} (\delta_{ij} + e_{ij})^2 \tag{7.71}$$

Die *Maximum Likelihood* (ML) Schätzmethode lässt sich näherungsweise durch folgende gewichtete Zielfunktion beschreiben (siehe MacCallum et al., 2007)

$$F_{ML}(\boldsymbol{\xi}) = \sum_{i,j} \frac{(s_{ij} - \omega_{ij})^2}{u_i^2 u_j^2} = \sum_{i,j} \frac{(\delta_{ij} + e_{ij})^2}{\theta_{ii}\theta_{jj}} \tag{7.72}$$

wobei $u_i^2 = \sigma_{ii} - \omega_{ii} = \theta_{ii}$ die Varianz des Itemresiduums ist. In der ML-Schätzung werden also Items in Itempaaren in der Zielfunktion hochgewichtet, die eine kleine Residualvarianz besitzen. Die Idee ist dabei, dass diese Items besonders informativ für die Schätzung der Itemladungen $\mathbf{\Lambda}$ und Kovarianzmatrix der Faktoren $\mathbf{\Phi}$ sind. Wenn keine Modellfehler δ_{ij} existieren, ist diese Gewichtung im Hinblick auf die Genauigkeit der Schätzer am effizientesten. Bei hohen Korrelationen s_{ij} und damit kleinen Residualvarianzen u_i^2 und u_j^2 entsteht ohne Existenz von Modellfehlern bei der ML-Anpassung nur ein kleiner Sampling-Fehler. Bei der ULS-Anpassung existiert keine explizite Verteilungsannahme für Residuen und Modellfehler (MacCallum et al., 2007).

Wenn jedoch (kleine) Modellfehler existieren (die schwache Faktoren abbilden), so zeigen MacCallum et al. (2007), dass die ULS-Schätzung robuster als die ML-Schätzung ist und zu weniger verzerrten Parameterschätzungen führt. Die ULS-Zielfunktion beruht implizit auf einer Normalverteilungsannahme von δ_{ij} und e_{ij}, wenn man die Zielfunktion als korrespondierende Likelihood auffasst. Betrachtet man Modellfehler δ_{ij} als nicht systematisch gemäß einer (hypothetischen) Normalverteilung um Null verteilt (etwa gemäß einer t-Verteilung oder einer Mischverteilung mit einzelnen Modellfehlern δ_{ij}, die „Ausreißer“ darstellen), dann kann man robustere Zielfunktionen zur Modellanpassung wählen, wie beispielsweise Siemsen und Bollen (2007)

$$F_{MAD}(\boldsymbol{\xi}) = \sum_{i,j} |s_{ij} - \omega_{ij}| = \sum_{i,j} |\delta_{ij} + e_{ij}| \tag{7.73}$$

Im Allgemeinen führen Modellfehler in Faktormodellen dazu, dass Itemladungen und Faktorkorrelationen nicht stabil sind, d.h. unter Weglassen einzelner Items oder Itemgruppen können sich die Ergebnisse ändern (siehe wiederum das Stabilitätskonzept von Michailidis & de Leeuw, 1998). Dies führt beispielsweise in konfirmatorischen Faktorenanalysen dazu, dass Itemladungen einer Skala A in einem eindimensionalen Faktormodell mit nur diesem Faktor A gegenüber einem zweidimensionalen Faktormodell (mit Einfachladungsstruktur) mit zwei Skalen A und B verschieden ausfallen. Die Eigenschaften von Messmodellen und damit die Bedeutung der latenten Variablen werden daher in konfirmatorischen Faktormodellen (und allgemeiner in Strukturgleichungsmodellen) nicht fixiert, so dass die Bedeutung der latenten Variablen über Modelle hinweg variiert.

Anhänger reflektiver Messmodelle werfen formativen Messmodellen häufig vor, dass die formative Definition latenter Variablen eine Abhängigkeit der Bedeutung der Variablen von anderen Variablen nach sich zieht (siehe z.B. Howell, 2014 oder Rhemtulla, Bork & Borsboom, 2015). Dieser Aussage ist zuzustimmen. Das gleiche Problem trifft aber auch für reflektive Messmodelle zu, denn in der Forschungspraxis der Strukturgleichungsmodellen (SEM) mit reflektiven definierten latenten Variablen werden die Messmodelle in jedem neu zu spezifizierenden SEM jeweils neu geschätzt.

Es sei an dieser Stelle betont, dass man nach unserer Meinung die Anpassung von Messmodellen (Schätzung von Itemparametern für einzelne latenten Variablen) von der Anpassung von Strukturmodellen (Beziehungen zwischen latenten Variablen untereinander und weiteren manifesten Variablen) trennen sollte (Anderson & Gerbing, 1988; McDonald, 2010; Williams & O'Boyle, 2011) und für die Analyse von Strukturmodellen nur Pfadmodelle mit Messfehlerkorrekturen (*single indicator models*) einsetzen sollte (Hayduk, 1987;

Oberski & Satorra, 2013). Alternativ könnten dazu Plausible Values als Realisierungen latenter Variablen verwendet werden (Mislevy, 1991; Asparouhov & Muthén, 2010).

Für einen festen Test mit I Items wird der Modellfehler in der (traditionellen) Forschungspraxis mit einer festen Größen des Modellmisfits (einer Modellfit-Statistik) aufgefasst. Eine Beurteilung von Faktormodellen hängt davon ab, ob bei der Anpassung eines Modells die Modellfehler „klein genug“ ausfallen. Ein bedeutsamer Modellfehler δ_{ij} kann dabei als nichtmodellierte Residualkovarianz θ_{ij} aufgefasst werden.

Im kommenden Abschnitt wird eine kürzlich vorgeschlagene stochastische Betrachtung von Modellfehlern diskutiert.

Stochastische Modellierung von Modellfehlern nach Wu und Browne

In einem jüngst erschienenen Ansatz diskutieren Wu und Browne (2015) eine stochastische Modellierung von Modellfehlern. Dabei wird von einem mehrdimensionalen Faktormodell mit einer festen Itemanzahl I ausgegangen. Modellfehler bezeichnen Wu und Browne (2015) dabei als *adventitious errors*. In einer Stichprobe mit n Personen liegt dabei eine empirische Kovarianzmatrix $\boldsymbol{S}$ der Dimension $I \times I$ vor. Die Matrix $\boldsymbol{S}$ ist Wishart-verteilt mit N Freiheitsgraden und einer Populationskovarianzmatrix $\boldsymbol{\Sigma}$. Die Populationskovarianzmatrix wird wiederum durch ein Faktormodell $\boldsymbol{\Sigma} \approx \boldsymbol{\Lambda\Phi\Lambda}^T + \boldsymbol{\Theta} =: \boldsymbol{\Omega} = \boldsymbol{\Omega}(\boldsymbol{\xi})$ repräsentiert. Es existiert dabei ein Modellfehler, d.h. die Matrix $\boldsymbol{\Delta} = \boldsymbol{\Sigma} - \boldsymbol{\Omega}$ ist verschieden von Null.

Die Stochastizität des Modellfehlers wird in Wu und Browne (2015) durch die Annahme modelliert, dass $\boldsymbol{\Sigma}^{-1}$ Wishart-verteilt mit der Skalenmatrix $\boldsymbol{\Omega}^{-1}$ und $m > I - 1$ Freiheitsgraden (dem sog. *Präzisionsparameter*). Dabei ist zu betonen, dass sowohl die modellimplizierte Kovarianzmatrix $\boldsymbol{\Omega}$ als auch der Präzisionsparameter m geschätzt werden. Die Abweichung zwischen der empirischen Kovarianzmatrix $\boldsymbol{S}$ und der Populationskovarianzmatrix $\boldsymbol{\Sigma}$ ist eine Funktion der Stichprobengröße N der Personen. Je mehr Personen vorliegen (d.h. $N \to \infty$), desto mehr nähert sich $\boldsymbol{S}$ der Matrix $\boldsymbol{\Sigma}$ an. In Analogie kann man sich den Parameter m als einen dem Modellfehler korrespondierenden „Stichprobenumfang“ vorstellen. Für ein sehr großes m stimmen $\boldsymbol{\Sigma}$ und $\boldsymbol{\Omega}$ praktisch überein, der Modellfehler verschwindet also. Ein relativ kleines m (z.B. $m = 50$) drückt bedeutsame Modellfehler $\boldsymbol{\Delta}$ aus.

Anstelle des Präzisionsparameters m verwenden Wu und Browne (2015) ebenso den „mittleren Modellfehler“ $\nu = 1/(m - I + 1)$. Satorra (2015) argumentiert, dass durch das von Wu und Browne (2015) vorgeschlagene Modell der Parameter $\boldsymbol{\xi}$ erwartungstreu geschätzt werden kann und außerdem $E(\boldsymbol{S}) = E(\boldsymbol{\Sigma}) = \boldsymbol{\Omega}$ gilt, d.h. trotz Modellfehler wird „im Mittel“ die richtige Kovarianzmatrix reproduziert.

Wu und Browne (2015) schlagen verschiedene Maximum Likelihood basierte Schätzmethoden für die Parameterschätzung vor. Sie leiten die asymptotische Verteilung der Diskrepanzmatrix $\boldsymbol{S} - \boldsymbol{\Omega}$ und des interessierenden geschätzten Parameters $\hat{\boldsymbol{\xi}}$ in Abweichung vom Populationsparameter $\boldsymbol{\xi}$ ab. Damit ist also eine simultane statistische Inferenz für Modellparameter unter Berücksichtigung des Samplingfehlers durch Personen und des Modellfehlers möglich. Im traditionellen Ansatz der Schätzung von Faktormodellen werden Standardfehler (fälschlicherweise) unter der Annahme eines exakt passenden Modells berechnet. Zusätzlich wird eine χ^2-Verteilung für die Schätzung $\hat{\nu}$ abgeleitet und es wird

gezeigt, dass die Beziehung $\sqrt{\hat{\nu}} = RMSEA$ praktisch erfüllt ist.

Zusammengefasst ist im Ansatz von Wu und Browne (2015) der Modellfehler in einem stochastischen Modell nur durch den Präzisionsparameter m parametrisiert. Dies kann eine strikte Verteilungsannahme sein, parametrisiert aber ggf. dadurch geeignet „zufällige Modellabweichungen" der Kovarianzmatrix $\boldsymbol{\Sigma}$ von $\boldsymbol{\Omega}$. Die Einbettung in den traditionellen Ansatz ist dadurch gegeben, dass sich der stochastisch definierte Modellfehler in der Größe $\nu = \nu(m)$ transformieren lässt, die in direkter Beziehung zum $RMSEA$ steht.

Für IRT-Modelle mit dichotomen oder polytomen Daten lässt sich die Modellierung von Wu und Browne (2015) direkt durch die Verwendung einer tetrachorischen oder polychorischen Korrelationsmatrix $\boldsymbol{S}$ einsetzen. Bei nur wenigen Items kann man auch die Approximation der Verteilung der Item Response Pattern $P(\boldsymbol{X})$ betrachten. Durch Sampling von Personen entsteht eine empirische Verteilung mit relativen Häufigkeiten $\hat{P}(\boldsymbol{X})$. Die Verteilung $P(\boldsymbol{X})$ in der Population wird dann durch eine Verteilung $Q_{\boldsymbol{\xi}}(\boldsymbol{X})$ approximiert, die von zu schätzenden Modellparametern $\boldsymbol{\xi}$ (Itemparametern und Parametern der Verteilung der latenten Variablen θ) abhängt. Der Modellfehler $\Delta(\boldsymbol{X}) = P(\boldsymbol{X}) - Q_{\boldsymbol{\xi}}(\boldsymbol{X})$ könnte durch ein Sampling von $P(\boldsymbol{X})$ aus einer Dirichlet-Verteilung $\mathcal{D}$ mit Anteilen $Q_{\boldsymbol{\xi}}(\boldsymbol{X}) = (q_1, \ldots, q_K)$ und einer zu schätzenden „Stichprobengröße" m erfolgen (siehe Gelman et al., 2004). Genauer wird dabei die Dirichlet-Verteilung $\mathcal{D}(m \cdot q_1, \ldots, m \cdot q_K)$ betrachtet, wobei der Präzisionsparameter m zu schätzen ist. Damit erhält man ein äquivalentes Vorgehen zu Wu und Browne (2015).

Approximate Factor Model und Domain Sampling

In den letzten beiden Abschnitten diskutierten wir Faktormodelle mit einer festen Anzahl von Items und haben gezeigt, dass Modellfehler stochastisch modelliert werden können. In der Perspektive des Domain Samplings existieren unendlich viele Items in der Population. Chamberlain und Rothschild (1983) gehen davon aus, dass ein Faktormodell niemals exakt gelten wird (vgl. auch MacCallum & Tucker, 1991). Ein konkreter Test mit I Items führt dabei zu einer beobachteten Kovarianzmatrix $\boldsymbol{S}_I$, die eine Populationskovarianzmatrix $\boldsymbol{\Sigma}_I$ schätzt. Auf Populationsebene wird dann für $\boldsymbol{\Sigma}_I$ ein Faktormodell $\boldsymbol{\Omega}_I := \boldsymbol{\Lambda}_I \boldsymbol{\Phi}_I \boldsymbol{\Lambda}_I^T + \boldsymbol{\Theta}_I$ postuliert. Im Allgemeinen wird dabei ein Modellfehler $\boldsymbol{\Delta}_I = \boldsymbol{\Sigma}_I - \boldsymbol{\Omega}_I$ existieren. D.h. das statistische Modell für jeden Test mit I Items und K Faktoren lautet

$$\boldsymbol{\Sigma}_I = \boldsymbol{\Lambda}_I \boldsymbol{\Phi}_I \boldsymbol{\Lambda}_I^T + \boldsymbol{\Theta}_I + \boldsymbol{\Delta}_I \tag{7.74}$$

Chamberlain und Rothschild (1983) bezeichnen dieses Modell als *approximate factor model*. Sie leiten dann Aussagen unter der Annahme ab, dass der $(K+1)$-größte Eigenwert der Kovarianzmatrix $\boldsymbol{\Sigma}_I$ für alle Testlängen I beschränkt ist. Dies sichert, dass asymptotisch nur K latente Variablen bedeutsam sind und der Modellfehler asymptotisch im Hinblick auf Parameterschätzungen an Bedeutung verliert. Man kann diese Eigenschaft mit Annahmen der Arbeiten zu essenziell eindimensionalen IRT-Modellen vergleichen (Stout, 1990; Junker, 1991).

Das Konzept des approximate factor model ist für konfirmatorische Faktormodelle und IRT-Modelle übertragbar. Neben dem Person Sampling führt auch das Item Sampling (vgl. Kapitel 4) zu einer Variabilität in resultierenden Parameterschätzungen. Damit sind Itemparameter (Itemladungen, Item Intercepts) und Faktorkovarianzen nur in der

Population von unendlich vielen Personen und unendlich vielen Items definiert. Beispielsweise kann eine geschätzte Itemladung $\hat{\lambda}_i$ in einem Test mit I Items eine Schätzung der Itemladung dieses Items bezüglich der in der gesamten Itemdomäne definierten Itemladung $\lambda_{i,\infty}$ sein. Im Allgemeinen wird sich die Itemladung $\lambda_{i,\infty}$ aber von der Itemladung $\lambda_{i,I}$ in der Population der Personen und dem festen Test mit I Items unterscheiden. Wenn das approximate factor model das zugrunde gelegte Testmodell sein soll, dann ist eine Modellfit-Statistik für den Test mit I Items im Hinblick auf die Beurteilung einer möglichen Verzerrung von Parameterschätzungen nicht aussagekräftig.

Für die simultane statistische Inferenz von Personen und Items unter der Person und Item Sampling Perspektive bietet sich wiederum der Double Jackknife (Brennan, 2001a, Kap. 6) Ansatz an, bei dem einzelne Personen (oder Personengruppen; etwa bei Clusterstrukturen der Schachtelung von Personen in Klassen) und Items (oder Itemgruppen; etwa bei in Testlets vorgegebenen Items) aus der statistischen Analyse entfernt werden. Wie im Ansatz von Wu und Browne (2015), die die Modelldiskrepanz mit dem Maß ν als (praktische) Transformation des $RMSEA$ beschreiben, kann man jedoch auch im Double Jackknife auf Modellresiduen beruhende „klassische Modellfit-Statistiken“ berechnen und die statistische Inferenz simultan für Personen und Items durchführen.

Für Cronbachs Alpha haben wir in Abschnitt 7.3.3 die statistische Inferenz im Hinblick auf Item Sampling mit einem Varianzkomponentenmodell abgebildet. Ein ähnliches Modell könnte auch für die Betrachtung einer latenten Kovarianz zweier Skalen X und Y eingesetzt werden. Sei dazu σ_{ij}^{XY} die Kovarianz von Item i (zugehörig zu Skala X) und Item j (zugehörig zu Skala j). Dann könnte ein Varianzkomponentenmodell definiert werden als

$$\sigma_{ij}^{XY} = \rho^{XY} + u_i^X + v_j^Y + \epsilon_{ij} \tag{7.75}$$

mit unkorrelierten Itemeffekten u_i^X und v_j^Y sowie unkorrelierten Residuen ϵ_{ij}. Dabei ist ρ^{XY} die latente Kovarianz der Skalen X und Y in der Population aller Items. Mit diesem Modell soll abgebildet werden, dass die Kovarianz zweier latenter Variablen in einem Faktormodell davon abhängig sein kann, welche Items als Indikatoren für die beiden Faktoren verwendet werden.

Für das Modell (7.75) können wiederum Kleinste-Quadrat-Schätzungen, Maximum-Likelihood-Schätzungen oder MCMC-Schätzungen abgeleitet werden. In Modell (7.75) könnten außerdem Itemladungen eingefügt werden, so dass man insgesamt den Ansatz

$$\sigma_{ij}^{XY} = \lambda_i^X \lambda_j^Y \rho^{XY} + u_i^X + v_j^Y + \epsilon_{ij} \tag{7.76}$$

erhält. Dabei sollte ein Normierungsbedingung an die Ladungen gestellt werden, z.B. der Mittelwert aller Itemladungen einer Skala beträgt Eins. Man könnte allerdings auch eine bivariate hierarchische Verteilung (λ_i^X, u_i) für die Itemeffekte annehmen (siehe Fox, 2010), in der man den Erwartungswert der Ladungen auf Eins setzt.

In der Ökonometrie werden sog. hochdimensionale Faktormodelle im Kontext von Paneldaten in den letzten Jahren verstärkt diskutiert (z.B. Arellano & Bonhomme, 2011; Bai & Ng, 2002; Bai, 2003; Bai & Ng, 2008; Bai & Li, 2012b; Fan, Fan & Lv, 2008; Fan, Liao & Liu, 2015; Fernández-Val & Vella, 2011; Onatski, 2012). Dabei liegen von N Einheiten (Firmen, Ländern) T Zeitpunkte vor. Statistische Inferenz wird dann häufig für den Fall $N \to \infty$ und $T \to \infty$ simultan durchgeführt. Die statistische Theorie dieser Modelle

lässt sich auf psychometrische Fragestellungen übertragen, da in der ökonometrischen Literatur auch Modelle für dichotome Paneldaten diskutiert werden. Generell nutzen die verwendeten statistischen Techniken der hochdimensionalen Faktormodelle teilweise die *random matrix theory*, die Theorie unendlichdimensionaler stochastischer Matrizen (für statistische Anwendungen: Bouchaud & Potters, 2009; Harding, 2012; Paul & Aue, 2014; für mathematische Grundlagen: Edelman & Rao, 2005; Bai & Silverstein, 2010).

Ich glaube, dass die Perspektive der statistischen Inferenz in den Sozialwissenschaften nicht nur im Hinblick auf $N \to \infty$ (unendlich viele Personen), sondern auch für $I \to \infty$ (unendlich viele Items) geschehen sollte. Gerade Konstrukte in der Kompetenzmessung scheinen sich aus unserer Sicht nicht auf die konkret in einem Test verwendete Itemmenge, sondern auf eine größere (hypothetische) Itempopulation zu beziehen.

Wenn man jedoch die Menge der Items in einem Test oder einem Fragebogen als fest ansieht, so sprechen auch statistische Gründe für den Einsatz von Verfahren, die die statistische Inferenz für $N \to \infty$ und $I \to \infty$ betrachten. Viele typischerweise eingesetzte Verfahren (wie Maximum Likelihood Schätzungen für Faktormodelle oder Strukturgleichungsmodelle) beruhen auf der Annahme, dass die Anzahl N der Personen groß gegenüber der (meist als fest betrachteten) Anzahl der Variablen I ist. Dies ist aber vor allem bei vielen kleiner angelegten sozialwissenschaftlichen Studien mit Fragebogendaten nicht der Fall (z.B. $N = 200$ Personen, für die $I = 100$ Variablen mit einem Fragebogen erhoben werden). Maximum Likelihood Schätzungen in Faktormodellen verwenden dabei die Inverse der modellimplizierten Kovarianzmatrix $\mathbf{\Sigma}^{-1}$ in der Zielfunktion. Werden viele Parameter in einem Faktormodell bei kleinem N geschätzt, so ist die Schätzung von $\mathbf{\Sigma}^{-1}$ sehr variabel, was zu großen Standardfehlern von Modellparametern führt. Zur stabilen Schätzung von Parametern in Faktormodellen scheinen dabei sog. Regularisierungsverfahren bedeutsam (siehe Yuan & Chan, 2008 oder Pourahmadi, 2013 für Faktormodelle und Bühlmann & Van De Geer, 2011 für weitere statistische Modelle). Bei kleinen Stichproben sollten daher mit dem Ziel stabiler Parameterschätzungen nur die „relevantesten“ Parameter geschätzt werden, „irrelevante“ Parameter werden mittels Regularisierungsverfahren auf Null oder „kleine Werte“ gesetzt (siehe auch Abschnitt 7.3.5 für eine Anwendung bei der DIF-Testung; für einen Überblick siehe Bickel & Li, 2006 oder Tutz, 2012).

Modellfehler und Domain Sampling in IRT-Modellen

Item-Response-Modelle werden häufig mit *full information maximum likelihood* Verfahren angepasst. Trifft man die Normalverteilungsannahme für latente Item Responses, so kann man sich anstelle der Anpassung des vollen Item Response Patterns (d.h. Zusammenhängen der I Items bis zur I-ten Ordnung) auf bivariate Zusammenhänge beschränken, die in tetrachorischen oder polychorischen Korrelationsmatrizen abgebildet sind (Takane & De Leeuw, 1987; Kamata & Bauer, 2008). Die obigen Überlegungen zu Faktormodellen und dem Domain Sampling sind damit direkt auf IRT-Modelle übertragbar. McDonald (1999, Kap. 12) argumentiert, dass für die Anpassung von IRT-Modellen empirisch kaum Informationen über die bivariaten Zusammenhänge hinaus benötigt werden (vgl. auch Maydeu-Olivares & Joe, 2014 im Kontext von Modellfit-Statistiken für IRT-Modelle).

In einem Strang der Literatur zur essenziellen Eindimensionalität (Stout, 1990; Junker, 1991, 1993; Ellis & Junker, 1997) werden Modellfehler explizit zugelassen. Diese Modell-

fehler stellen Abweichungen von der lokalen stochastischen Unabhängigkeit dar. Ein IRT-Modell mit einer eindimensionalen latenten Variablen und monotonen Item-Response-Funktionen wird in diesen Arbeiten durch Bedingungen in der Domäne mit unendlich vielen Items charakterisiert, so dass nichtmodellierte Abhängigkeiten sich nicht in relevanten weiteren Dimensionen niederschlagen (vgl. dazu die Bedingungen für approximate factor models). Es wird gezeigt, dass Personenfähigkeiten konsistent im Hinblick auf Item Sampling (d.h. für $I \to \infty$) unter der Modellannahme essenzieller Eindimensionalität geschätzt werden können.

Wir hatten in Abschnitt 5.2 in den vorangehenden Teilen dieses Abschnittes diskutiert, dass die Annahme der lokalen stochastischen Unabhängigkeit der Annahme unkorrelierter Residuen in Faktormodellen entspricht. In Faktormodellen bei einer Anpassung mit unweighted least squares (ULS) mitteln sich daher die Modellfehler notwendigerweise aus (siehe Abschnitt 5.2). Die Annahme, dass alle Residualkorrelationen gleich Null sind, wird also bei der Modellanpassung durch die schwächere Annahme ersetzt, dass der Mittelwert der Residualkorrelationen gleich Null ist. Allerdings ist diese Annahme ähnlich zum in Wu und Browne (2015) diskutierten Modellfehler: „Im Mittel" werden dabei jeweils die richtigen Kovarianzmatrizen reproduziert.

Die Konsistenz von Personenparameterschätzungen für $I \to \infty$ (unendlich viele Items) leiten Clarke und Junker (1991) her. Dabei gehen sie von Verletzungen der lokalen stochastischen Unabhängigkeit aus. In einem ersten Ansatz wird anstelle der „wahren Likelihood" $\nu(\boldsymbol{X}_p|\theta)$ mit lokalen Abhängigkeiten eine fehlspezifizierte Likelihood $\tilde{p}(\boldsymbol{X}|\theta) = \prod_i p(X_{pi}|\theta)$ eingesetzt. Bezüglich der Kullback-Leibler-Information ist $\tilde{p}$ die Bestapproximation von ν (Clarke & Junker, 1991; vgl. auch Haberman, 2007). Man kann sich diese Approximation analog zur linearen Regression vorstellen, wobei „im Mittel" das Regressionsmodell die „richtigen Aussagen" trifft. Die Schätzung mit einer misspezifizierten Likelihood wird auch als *Quasi Maximum Likelihood* bezeichnet (White, 1982). Unter schwachen Annahmen wird dann die Konsistenz der Personenparameterschätzung und die asymptotische Normalverteiltheit gezeigt. In einem zweiten Ansatz wird diese Eigenschaft (unter gewissen Annahmen) ebenso erhalten. Dabei wird angenommen, dass die wahre Likelihood neben einer primär interessierenden Dimension θ noch weitere Störndimensionen (*nuissance dimensions*) besitzt und die Modellanpassung diese Störndimensionen ignoriert (Clarke & Junker, 1991). Es werden Bedingungen aufgestellt, die sichern, dass das Ausintegrieren der Störndimensionen „keinen Einfluss" auf die interessierende Dimension θ besitzt (vgl. auch Ip, 2010).

Die Aufgabe der Forderung lokaler stochastischer Unabhängigkeit (und damit das explizite Zulassen von Modellfehlern) wird in der Replik von Eckes (2015b) auf den Kommentar von Robitzsch und Lüdtke (2015) stark kritisiert, sie erscheinen „außerordentlich problematisch"[18]. Eckes (2015b) merkt an, dass nur postuliert werden würde, dass sich „positive und negative lokale Abhängigkeiten gegenseitig aufheben könnten". Im Gegenteil, durch die ULS-Schätzung ist (wie in der linearen Regression) gesichert, dass die Summe der Modellfehler (d.h. der Regressionsresiduen) gleich Null ist. Weiter wird „von den nachteiligen Folgen für die Schätzung von Personen- und Itemparametern" (Eckes, 2015b) gesprochen. Diese Aussage kann nur metaphorisch gemeint sein, denn es ist unklar,

[18]Damit erfolgt die Kritik von Eckes (2015b) in einer ähnlichen Richtung wie die von Steyer, Sengewald und Hahn (2015) auf Wu und Browne (2015).

auf welche Modellannahme sich diese Kritik überhaupt bezieht. Man kann mit der ULS-Schätzung unter Annahme von Modellfehlern oder dem stochastischen Ansatz von Wu und Browne (2015) zu konsistenten Item- und Personenparameterschätzungen gelangen. Conditional maximum likelihood Schätzungen und marginal maximum likelihood Schätzungen für IRT-Modelle teilen jedoch auch nur diese Eigenschaft, so dass für endlich viele Personen und/oder endlich viele Items keine unverzerrte Parameterschätzung gesichert ist (vgl. auch van der Linden, 1994). Damit bleibt unklar, weshalb die Schätzungen so „nachteilige" Eigenschaften haben sollten. Die nichtmodellierte Mehrdimensionalität (d.h. der Modellfehler) würde außerdem die „Validität und Fairness eines Tests [...] gefährden" (Eckes, 2015b). Hier wird der Standpunkt eingenommen, dass die Validität nicht zwingend durch die Passung eines psychometrischen Modells gezeigt werden muss. Der Begriff der „Fairness" wird zwar häufiger im Kontext von Rasch-Modellierungen verwendet, um invariantes Itemfunktionieren in allen (relevanten) Subpopulationen zu fordern, scheint für uns aber eher historische als aktuelle Bedeutung im Educational Assessment zu besitzen (siehe Camilli, 2006; vgl. auch eine entsprechende Argumentation in Robitzsch et al., 2015). Fairness sollte nach unserer Meinung eher an Außenkriterien orientiert sein. Diese Kritikpunkte finden sich aber auch in ähnlicher Weise für die stochastische Modellierung von Modellfehlern nach Wu und Browne (2015) wieder, die in einer Aufgabe der „klassischen Bedeutung" latenter Variablen münden würde (Steyer et al., 2015). Vielmehr scheinen latente Variablen in dieser Interpretation nur für einen festen Test mit I Items und unter der Annahme lokaler stochastischer Unabhängigkeit (oder einer anderen Identifikationsannahme, die zu einem passenden psychometrischen Modell führt) wohldefiniert. Unter der Domain Sampling sind für uns latente Variablen erst in der gesamten Itempopulation definiert, d.h. durch die Annahme eines statistischen Modells, das die Inferenz $I \to \infty$ beinhaltet.

Praktisch kann die simultane statistische Inferenz für Personen und Items wie für Faktormodelle wiederum mit der Methode des Double Jackknife (Brennan, 2001a, Kap. 6) durchgeführt werden. Itemschwierigkeiten, Itemladungen und Modellfit-Statistiken können dabei in Abhängigkeit von der Itemauswahl variieren. Eher auf illustrativer Ebene wurde die Abhängigkeit von Itemparametern für Bifaktor-Modelle unter Weglassen einzelner Subskalen von Reise (2012, S. 683ff.) untersucht und er findet, dass die Itemladungen auf dem g-Faktor in Abhängigkeit von den verwendeten Items verschieden ausfallen. Bei einer großen Stichprobe von Personen kann man also entweder von einem Modellfehler bei einem Interpretation einer festen Itemmenge oder von „zufälliger Variabilität" aufgrund eines Domain Samplings ausgehen. Dieses Beispiel verdeutlicht, dass absolute Maße des globalen Modellfits vielleicht weniger hilfreich sind, wenn der Fokus auf einzelnen Modellparametern liegt. Für diese Parameter ist demzufolge die Variabilität in Abhängigkeit von Person und Item Sampling zu studieren (vgl. dies wiederum mit dem Ansatz der *Replikationsstabilität* von Michailidis & de Leeuw, 1998; siehe auch Gifi, 1990).

Bayesianische Faktormodelle

In jüngerer Zeit werden verstärkt Bayesianische Ansätze zur Schätzung von Faktormodellen diskutiert (Kaplan, 2014; Lee, 2007; Song & Lee, 2012). Neben der immer stärker verwendeten Software WinBUGS (Lunn et al., 2000) ist das Interesse an Bayesiani-

schen Methoden mit der Implementierung dieser Methodik in Mplus (Muthén & Muthén, 1998-2013) gewachsen. Konzeptuell werden in der Bayesianischen Statistik unbekannte Parameter γ als Zufallsvariable und nicht als fester, unbekannter Wert wie in der frequentistischen Statistik angesehen. Demzufolge besitzt ein Parameter γ eine Priorverteilung $g(\gamma)$, die „Vorwissen“ für den Parameter charakterisiert. Für die Anpassung eines statistischen Modells liegen Daten $\boldsymbol{X}$ vor, aus der die Likelihood $L(\boldsymbol{X}|\gamma)$ gebildet wird. Das Maximum der Likelihood führt zur Likelihood-Schätzung von γ. Gemäß der Formel von Bayes (Gelman et al., 2004) ergibt sich die Posteriorverteilung h in der Bayes-Statistik gemäß

$$h(\gamma|\boldsymbol{X}) = \frac{L(\boldsymbol{X}|\gamma) \cdot g(\gamma)}{\int L(\boldsymbol{X}|\gamma) \cdot g(\gamma) \mathrm{d}\gamma} \propto L(\boldsymbol{X}|\gamma) \cdot g(\gamma) \tag{7.77}$$

wobei das Symbol „$\propto$“ andeuten soll, dass Gleichheit bis auf eine Multiplikationskonstante gegeben ist. In der Bayes-Statistik kann der Maximalwert der Posteriorverteilung h als Schätzer für γ verwendet werden (Gelman et al., 2004; siehe auch Kaplan, 2014). Logarithmiert man (7.77) und schreibt $l = \log L$, so erhält man

$$\log h(\gamma|\boldsymbol{X}) \propto l(\boldsymbol{X}|\gamma) + \log g(\gamma) \tag{7.78}$$

Man sieht anhand von (7.78), dass die Maximierung auf zwei Informationen beruht: die Information aus den Daten $l(\boldsymbol{X}|\gamma)$ und die Information g aus der Priorverteilung. Ist die Priorverteilung g näherungsweise konstant, d.h. $g(\gamma) = C$, dann stimmen die Maximum Likelihood und die Bayes-Schätzung überein. Ohne die Interpretation von γ als Zufallsvariable kann man aber (7.78) als Optimierung unter einer modifizierten Likelihood ansehen. Typischerweise wird dadurch die Likelihood *regularisiert* (man verwendet auch den Begriff *Penalized Maximum Likelihood* Schätzung; siehe z.B. DeCarlo, Kim & Johnson, 2011 für eine Anwendung bei IRT-Modellen). Ich schließe mich dieser Modellperspektive an und sehen die Verwendung von Priorverteilungen als pragmatische Möglichkeit, Schätzprobleme bei schwachen Datenlagen zu lösen und wünschenswerte frequentistische Eigenschaften von Schätzern zu erhalten.

Für konfirmatorische Faktorenanalysen diskutieren Muthén und Asparouhov (2012) Bayesianische Schätzmethoden, die in Mplus implementiert sind. Dabei monieren sie häufig vorzufindenden Modellmisfit bei konfirmatorischen Faktorenanaysen mit Einfachladungsstruktur. Muthén und Asparouhov (2012) schlagen informative (und fixierte) Priorverteilungen für Nebenladungen in $\boldsymbol{\Lambda}$ und die Residualkovarianzmatrix $\boldsymbol{\Theta}$ vor. Durch die Verwendung einer Priorverteilung für $\boldsymbol{\Theta}$ werden Modellfehler verringert. „Große Modellfehler“ dominieren dann auch die informative Priorverteilung und führen zu einer praktisch modellierten Residualvarianz, während „kleine Modellfehler“ von der informativen Priorverteilung dominiert werden.

Eine hierarchische Priorverteilung für die Residualkovarianz wird in Rowe (2002, Kap. 9) und Press (2003, Kap. 15) diskutiert. Dadurch können Modellfehler mit deutlich weniger Restriktionen als in Muthén und Asparouhov (2012) abgebildet werden.

Regularisierungsmethoden für die Schätzung der Residualkovarianzmatrix werden in Bai und Liao (2013) eingesetzt. Neben dem Term der Log-Likelihood wird dabei für die Schätzung ein Regularisierungsterm (*penalty*) $Pen(\boldsymbol{\Theta}) = \iota \cdot \sum_{i,j} |\theta_{ij}|$ mit einem zu schätzenden Regularisierungsparameter ι eingesetzt (vgl. auch Fan, Liao & Mincheva, 2013).

Der Term korrespondiert mit einer robusten Zielfunktion (vgl. die lineare Regression) und einer Priorverteilung, die wenige sehr große Werte und viele sehr kleine Werte besitzt. In der Schätzmethode werden die meisten Einträge in $\mathbf{\Theta}$ gleich Null geschätzt und nur die bedeutsamen Modellfehler erscheinen als Nicht-Nulleinträge in $\mathbf{\Theta}$. Die statistische Inferenz wird dabei simultan für $N \to \infty$ und $I \to \infty$ durchgeführt. Für eine Korrespondenz von Regularisierungsmethoden und einer speziellen Wahl von Priorverteilungen für lineare Regressionen siehe Park und Casella (2008).

Gegenüber Maximum Likelihood (ML) basierten Ansätzen lassen sich Modellfehler als Nichtdiagonaleinträge in der Residualkovarianzmatrix $\mathbf{\Theta}$ relativ einfach mit Bayesianischen Verfahren (MCMC) spezifizieren und schätzen. Es lässt sich jedoch feststellen, dass gegenüber ML-Ansätzen Modellfit-Statistiken für Bayes-Ansätze bei der Schätzung von Faktormodellen noch nicht abschließend ausgearbeitet sind (siehe jedoch Levy, 2011).

7.3.5 Bedeutung der Invarianztestung für Gruppenvergleiche

In diesem Abschnitt diskutieren wir die Bedeutung der Domain Sampling Perspektive für Differential Item Functioning (DIF; Holland & Wainer, 1993) bzw. für die Invarianztestung in Faktormodellen (Mellenbergh, 1989; Meredith, 1993; Millsap, 2011). Wir verwenden beide Begriffe im Folgenden synonym zueinander. Wir diskutieren DIF im Rasch-Modell und im 2PL-Modell (Faktormodell) separat. Danach übertragen wir die Überlegungen auf Längsschnittanalysen und des Linking in Large Scale Assessments.

Differential Item Functioning im Rasch-Modell

Wir diskutieren unsere Überlegungen für das Rasch-Modell für den Fall von zwei Gruppen $g = 1$ und $g = 2$. Dabei wird eine Kontrastvariable Z_{pg} eingeführt, die den Wert $1/2$ für Person p in Gruppe $g = 2$ und den Wert $-1/2$ für Person p in Gruppe $g = 1$ annimmt. Das Rasch-Modell mit differenziellen Itemfunktionieren kann dann geschrieben werden als

$$\text{logit}\, P(X_{pi} = 1|\theta) = \theta_p - b_i - Z_{pg}\delta_i \tag{7.79}$$

Der Parameter δ_i ist dann der DIF-Effekt und quantifiziert den relativen Schwierigkeitsunterschied für Schüler beider Gruppen für das i-te Item. Man muss anmerken, dass man für die Schätzung von Modell (7.79) eine Identifikationsbedingung benötigt. Für die Schätzung von Gruppenmittelwerten und damit der Mittelwertdifferenz $\Delta\mu$ zwischen beiden Gruppen muss man annehmen, dass sich die DIF-Effekte δ_i zu Null summieren, d.h. $\sum_{i=1}^{I} \delta_i = 0$. Die Größe $\sigma^2_{DIF} = \sum_i \delta_i^2$ wird auch als *DIF Varianz* (Longford et al., 1993) bezeichnet.

Wenn die DIF-Effekte alle gleich Null sind, so liegt eine sog. Invarianz vor. Dies bedeutet, dass bei jedem der I Items der mittlere Unterschied zwischen beiden Gruppen homogen gemessen wird. Mitunter wird argumentiert, dass bei der Existenz von DIF-Effekten, „Testfairness“ nicht gegeben sei, denn beobachtete Unterschiede zwischen Gruppen lassen sich nicht auf „wahre Unterschiede“ in θ (der latenten Variablen) zurückführen (z.B. Strobl, 2010). Ich erachte diese Definition für zirkulär: wie soll eine latente Variable für einen Test mit I Items unabhängig von diesen Items definiert sein? Für einen als fest betrachteten Test kann man argumentieren, dass die Mittelwertdifferenz zwischen beiden Gruppen

eindeutig durch die Bedingung $\sum_{i=1}^{I} \delta_i = 0$ identifiziert ist. Klarerweise ändert sich die Mittelwertdifferenz zwischen beiden Gruppen bei jeder Teilstichprobe von Items. Wenn der Test jedoch als fest betrachtet wird, sehe ich dies nicht als konzeptuelles Problem.

Aus der Perspektive der Replikationsstabilität (Michailidis & de Leeuw, 1998) kann man die Variabilität der Mittelwertdifferenz unter der Itemauswahl studieren. Führt man Jackknife unter Elimination einzelner Items durch, so beträgt die durch Itemauswahl bedingte Variabilität der Mittelwertdifferenz $\Delta\mu$ für den Standardfehler $SE(\Delta\mu) = \sigma_{DIF}/\sqrt{I}$. Bei einer festgehaltenen DIF-Varianz σ^2_{DIF} sinkt mit einer zunehmenden Anzahl von Items demzufolge der Standardfehler der Mittelwertdifferenz.

Mitunter nutzt man DIF-Analysen, um Items mit einem betraglich zu großen DIF ($|\delta_i|$) für die Untersuchung von Mittelwertdifferenzen zu entfernen. Dabei beruht die Argumentationslogik darauf, dass man dann den Gruppenvergleich nur auf „ fairen“ Items mit kleinen $|\delta_i|$-Werten beruhen lässt, denn diese messen die Differenz in einem homogenen Ausmaß. Dies bedeutet jedoch, dass das Konstrukt (und damit die Bedeutung von θ) umdefiniert wird, um eine „Passung“ mit dem Rasch-Modell zu erreichen (vgl. auch Robitzsch et al., 2015 für eine Kritik zum Ausschluss von DIF-Items bei Kompetenztests). Praktisch wird durch den Ausschluss von DIF-Items die Verteilung der originalen δ_i-Werte gestutzt, denn betraglich zu große Werte werden entfernt. Man könnte die DIF-Effekte δ_i jedoch auch als Modellfehler (siehe Abschnitt 7.3.4) verstehen, die stochastisch modelliert werden. Dabei können verschiedene hierarchische Verteilungen für DIF-Effekte spezifiziert werden. Fox (2010) spezifiziert eine hierarchische Normalverteilung mit einer zu schätzenden Varianz. De Boeck (2008) spezifiziert eine Mischverteilung (*random item mixture* model), für die angenommen wird, dass ein Teil der Items keine DIF-Effekte aufweist ($\delta_i = 0$) und für den anderen Teil der Items DIF-Effekte gemäß einer Normalverteilung modelliert werden (siehe auch De Boeck, Cho & Wilson, 2011). Mit der logistischen Lasso-Regression als Regularisierungsverfahren wird die Annahme abgebildet, dass die meisten der Items keinen DIF und nur wenige Items substanziellen DIF besitzen (Tutz & Schauberger, 2015; Magis, Tuerlinckx & De Boeck, 2015). Dieser Ansatz umgeht das Problem des multiplen Testens auf DIF-Effekte aller Items und identifiziert datengetrieben mit einer wohldefinierten Zielfunktion die Mittelwertdifferenz auf Basis der Items ohne DIF.

Identifiziert man Gruppenunterschiede mit der Bedingung der Summennormierung $\sum_{i=1}^{I} \delta_i = 0$, so lässt man gruppenspezifische – und damit ggf. bedeutsame DIF-Effekte – zu. Eine Anpassung des Rasch-Modells mit gemeinsamen Itemparametern ohne explizite Modellierung des DIF führt notwendigerweise zu einem schlechter passenden Modell. Man kann aber argumentieren, dass das Modell ohne DIF-Effekte einen Modellfehler für die Item Intercepts beinhaltet. Durch das Vorhandensein des Modellfehlers gehen aber die DIF-Effekte im Rasch-Modell in das Residuum über, so dass im Modell ohne DIF-Effekte im Allgemeinen eine kleinere extrahierte Traitvarianz als im Modell mit DIF-Effekten resultiert (siehe Abschnitt 7.1.1).

Unter der Domain Sampling Perspektive mit einer unendlich großen Itempopulation ist die Mittelwertdifferenz $\Delta\mu_\infty$ auf Basis aller Items der Population definiert. Für einen konkreten Test mit I Items wird man daher nicht davon ausgehen, dass selbst bei einer unendlich großen Personenstichprobe die ermittelte Mittelwertdifferenz $\Delta\mu_I$ mit der Populationsgröße $\Delta\mu_\infty$ übereinstimmt. Ich erachte die Annahme, dass man beispielswei-

se für Geschlechterdifferenzen in der Mathematikkompetenz in allen Items nur homogene Unterschiede zulässt, für unplausibel (siehe Robitzsch et al., 2015). Natürlich gibt es gruppenspezifische relative Stärken und Schwächen (charakterisiert durch die DIF-Effekte δ_i), die im Allgemeinen konstruktinhärent sind (Roussos & Stout, 1996; siehe auch Schwabe & Gebauer, 2013) und mit einer höheren Testvalidität als bei einem rein statistisch getriebenen Ausschluss von DIF-Items einhergehen. Für Roussos und Stout (1996) wird DIF durch Gruppenunterschiede auf konstruktrelevanten Sekundärdimensionen verursacht. McCrae (2015) argumentiert, dass DIF durch Gruppenunterschiede in itemspezifischen Residuen verursacht werden kann. Da für ihn spezifische Varianz auch ein Bestandteil der wahren Varianz sein kann, sieht er Invarianz nicht als notwendige Voraussetzung für Gruppenvergleiche an. Für konträre Perspektiven siehe jedoch beispielsweise Little (2013) oder Molenaar und Borsboom (2013).

Das Ausmaß der Bedeutung der Itemauswahl (Item Sampling) für die Mittelwertdifferenz sollte in Publikationen mit Standardfehlern erfolgen, die das Person Sampling und das Item Sampling (bzw. die Stabilität unter Itemauswahl) separat ausweisen[19]. Computational kann der Ansatz wiederum mit der Double Jackknife Methode umgesetzt werden (Brennan, 2001a, Kap. 6).

Invarianz in Faktormodellen und Invariance Alignment

Im Folgenden diskutieren wir den Fall der Invarianz für mehrere Gruppen in eindimensionalen Faktormodellen bzw. im eindimensionalen 2PL-Modell der Item-Response-Theorie. Mit der Schreibweise latenter Item Responses X_{pgi} für Person p in Gruppe g und Item i ist

$$X^*_{pi} = \lambda_{ig}\theta_p + \nu_{ig} + \epsilon_{pgi} \tag{7.80}$$

Dabei verwenden wir die in den Faktormodellen übliche Schreibweise für Itemladungen λ und Item Intercepts ν.

In der Literatur konfirmatorischer Faktorenanalysen hat sich die Typologie verschiedener Grade der Invarianz nach Meredith (1993) eingebürgert (siehe auch Millsap, 2011). Man spricht von *konfiguraler Invarianz*, wenn in allen Gruppen g das eindimensionale Faktormodell gilt. *Schwache Invarianz* gilt, falls die Itemladungen λ_{ig} über Gruppen hinweg gleich sind. Sind zusätzlich auch die Item Intercepts ν_{ig} gleich über alle Gruppen, so spricht man von *starker Invarianz*. Ein Teil der Literatur zu Strukturgleichungsmodellen argumentiert, dass die Prüfung der Invarianz eine Voraussetzung für den Einsatz bestimmter statistischer Modelle sei (z.B. Beaujean, 2014). Für den Vergleich von Kovarianzen oder Regressionskoeffizienten mehrerer Gruppen sei dann schwache Invarianz nachzuweisen, für den Gruppenvergleich von Mittelwerten müsse starke Invarianz gelten (Beaujean, 2014, Kap. 4). In der Forschungspraxis betrachtet man noch „Zwischenstufen“ der sog. *partiellen Invarianz*, bei der einige Ladungen oder Intercepts zwischen Gruppen variieren dürfen. Bei diesem Ansatz scheint generell unklar, worauf sich die globale Zielfunktion bei der Modellanpassung bezieht. Damit ist gemeint, dass man in einem mehrstufigen Prozess eine Modellselektion betreibt. Dadurch wird im Allgemeinen die statistische Inferenz im

[19] Modellgeltungstests wie der Likelihood-Ratio-Test nach Anderson oder Modellvergleiche unter Annahme von invarianten und nichtinvarianten Items halte ich für nicht notwendig, um die Analyse von Mittelwertunterschieden zu legitimieren.

finalen Analysemodell inkorrekt sein, da sich diese Inferenz nur auf das betrachtete Modell bezieht und die Unsicherheit in den vorhergehenden Schritten ignoriert wird (vgl. auch Berk et al., 2010).

Ich argumentiere im Folgenden, dass Invarianz *keine* notwendige Voraussetzung von statistischen Mittelwerten darstellt.

In der Item-Response-Theorie stellt man eine Vergleichbarkeit von Gruppen (oder Studien) dadurch her, dass man unter einer Identifikationsbedingung für (7.80) (häufig mit $E(\theta) = 0$ und $Var(\theta) = 1$) zunächst in jeder der Gruppen unabhängig kalibriert (d.h. ein Faktormodell anpasst) und damit Itemparameter gewinnt. Durch ein anschließendes *Linking* (Kolen, 2006) werden die Itemparameter auf eine Metrik gebracht und Unterschiede zwischen Gruppen für Mittelwerte und Varianzen können bestimmt werden.

In dem Linking-Verfahren nach Haberman (2009) kann ein simultanes Linking für alle Gruppen mit Hilfe linearer Regressionen durchgeführt werden[20]. Dazu werden die Itemladungen zunächst logarithmiert, d.h. $l_{ig} = \log \lambda_{ig}$. Für diese Itemladungen wird ein Regressionsmodell

$$l_{ig} = \tilde{\sigma}_g + l_i + \epsilon_{l,ig} \tag{7.81}$$

postuliert. Als Modellparameter sind gemeinsame logarithmierte Itemladungen $l_i = \log \lambda_i$ sowie die logarithmierte Standardabweichung $\tilde{\sigma}_g = \log \sigma_g$ zu schätzen. Die Gleichung (7.81) lässt auch als

$$\lambda_{ig} = \exp(\tilde{\sigma}_g) \cdot \exp(l_i) \cdot \exp(\epsilon_{l,ig}) = \sigma_g \cdot \lambda_i \cdot \exp(\epsilon_{l,ig}) \tag{7.82}$$

schreiben und formalisiert die natürliche Forderung $\lambda_{ig} \approx \sigma_g \cdot \lambda_i$. Gilt für jede Gruppe g für die Fehler $\sum_i \epsilon_{l,ig} = 0$, so folgt für mit der Exponentialfunktion transformierten Fehler $\prod_i \exp(\epsilon_{l,ig}) = 1$. Definiert man dann die gelinkten gruppenspezifischen Itemparameter durch $\lambda^*_{ig} = \lambda_{ig}/\sigma_g$, so folgt $\prod_i \lambda^*_{ig} = \prod_i \lambda_i$, d.h. das Produkt der Itemladungen ist in allen Gruppen identisch, so dass man Unterschiede in Standardabweichungen zwischen den Gruppen identifizieren kann.

Die Modellparameter in (7.81) können durch eine Minimierung der kleinsten Quadrate geschätzt werden (siehe Haberman, 2009), so dass sich als Zielfunktion

$$F_{l,LS} = \sum_{i,g} \left(l_{ig} - \tilde{\sigma}_g - l_i\right)^2 \tag{7.83}$$

ergibt. Dies entspricht implizit der Annahme, dass die Residuen $\epsilon_{l,ig}$ als Abweichung von der Invarianz der Itemladungen normalverteilt sind. Sind andere Verteilungen plausibel oder existieren „Ausreißer“, so könnten wie im Fall der Regression robustere Zielfunktionen verwendet werden (Fox, 1997, Kap. 17). Die Standardabweichung von Gruppe g wird durch $\sigma_g = \exp(\tilde{\sigma}_g)$ bestimmt. In Analogie zur linearen Regression kann dann für die Regression (7.83) ein R^2_λ der Varianzaufklärung bestimmt werden. Invarianz bezüglich der Itemladungen gilt, falls R^2_λ gleich 1 ist. In einem zweiten Schritt wird das Linking der Item Intercepts vorgenommen. Mit den geschätzten Standardabweichungen σ_g aus

[20] Das Linking-Verfahren nach Haberman (2009) ist im R-Paket sirt (Robitzsch, 2015) in der Funktion `linking.haberman` implementiert.

dem ersten Schritt bildet man transformierte Item Intercepts $v_{ig} = \nu_{ig}/\lambda_{ig}$ [21]. Für diese Item Intercepts wird wiederum ein lineares Regressionsmodell $v_{ig} = v_i + \mu_g + \epsilon_{v,ig}$ postuliert und mit der Methode der kleinsten Quadrate die Parameterschätzungen für den Gruppenmittelwert μ_g und die Itemparameter v_i bestimmt. Wiederum kann man ein R^2_ν als Maß der Invarianz der Item Intercepts definieren. Für das Linking-Verfahren ist wiederum eine Identifikationsbedingung notwendig, etwa $\prod_g \sigma_g^2 = 1$ und $\sum_g \mu_g = 0$. Die Linking-Methode nach Haberman (2007) ist somit eine Verallgemeinerung des Mean-Sigma-Linkings für zwei Gruppen (siehe Kolen & Brennan, 2004).

Linking-Methoden lassen also Abweichungen von Invarianz zu und nehmen „nur“ an, dass sich diese Abweichungen ausmitteln. Für den Fall eines festen Tests kann man argumentieren, dass man nur irgendeine Identifikationsannahme benötigt, um Gruppenvergleiche hinsichtlich Mittelwerten und Varianzen durchzuführen. Unter der Domain Sampling Perspektive nimmt man an, dass „wahre Gruppenunterschiede“ im Fall unendlich vieler Items erhalten werden, jede Itemauswahl aus der Itemdomäne führt daher zu verschiedenen Gruppenunterschieden.

Einen ähnlichen Ansatz des Linkings von Itemladungen und Item Intercepts für Gruppenvergleiche verfolgen Asparouhov und Muthén (2014) (siehe auch Muthén & Asparouhov, 2014). In ihrem sog. *invariance alignment* (implementiert in der Software Mplus; Muthén & Muthén, 1998-2013) gehen sie zunächst wiederum davon aus, dass in den jeder der Gruppen unabhängig eindimensionale Faktormodelle geschätzt werden. Im Unterschied zu Haberman (2009) schlagen sie eine alternative Linking-Methode vor, die auf der paarweisen Auswertung von Unterschieden von Parametern zwischen allen Gruppen beruht. Asparouhov und Muthén (2014) minimieren folgende Zielfunktion simultan für Itemladungen und Item Intercepts[22]

$$F = \sum_{g,g',i} f\left(\frac{\lambda_{ig}}{\sigma_g} - \frac{\lambda_{ig'}}{\sigma_{g'}}\right) + \sum_{g,g',i} f\left(\nu_{ig} - \mu_g\frac{\lambda_{ig}}{\sigma_g} - \nu_{ig'} + \mu_{g'}\frac{\lambda_{ig'}}{\sigma_{g'}}\right) \tag{7.84}$$

In diesem Modell werden wiederum die Gruppenmittelwerte μ_g und Gruppenstandardabweichungen σ_g geschätzt (wieder mit einer Identifikationsbedingung wie im Linking nach Haberman). Dabei schlagen sie die Funktion $f(x) = \sqrt{\sqrt{x^2 + \varepsilon}}$ mit einem kleinen ε (z.B. $\varepsilon = .0001$) vor. Näherungsweise ist dann $f(x) = \sqrt{|x|}$, man erhält also eine „robuste“ Funktion. Dabei ist $f(x) > |x|$ für $|x| < 1$, kleine Abweichungen werden also stärker „bestraft“ als mit der Betragsfunktion sowie $f(x) < |x|$ für $|x| > 1$. Große Abweichungen werden also weniger stark bestraft (siehe Asparouhov & Muthén, 2014). Wenn die

[21] Es ist $X^*_{pgi} = \lambda_{ig}\theta_{pg} + \nu_{ig} + \epsilon_{pgi}$. Bei invarianten Itemladungen folgt aus (7.81) die Beziehung $\lambda_{ig} = \lambda_i\sigma_g$. Wir notieren weiter $\lambda_{ig}\theta_{pg} + \nu_{ig} = \lambda_{ig}(\theta_{pg} + \nu_{ig}/\lambda_{ig}) = \lambda_i\sigma_g(\theta_g + \mu_g + v_i)$, so dass man daraus $\theta_g \sim N(\mu_g, \sigma_g^2)$ erhält.

[22] Wir wollen die Motivation der Zielfunktion in (7.84) motivieren. Für die vorliegenden Itemparameter λ_{ig} und ν_{ig} wurde eine Standardnormalverteilung für θ in der Gruppe angenommen. Bezeichnen wir nun mit λ^*_{ig} und ν^*_{ig} die Itemparameter nach dem Linking, so ergibt sich für die Varianz $Var(X^*_{pgi}) = \lambda^2_{ig} = \left(\lambda^*_{ig}\right)^2\sigma_g^2$ und damit $\lambda^*_{ig} = \lambda_{ig}/\sigma_g$. Dieser Koeffizient wird nun für die Gruppen g und g' gleichgesetzt, d.h. $\lambda^*_{ig} = \lambda^*_{ig'}$.

Für den Erwartungswert ergibt sich $E(X^*_{pgi}) = \nu_{ig} = \lambda^*_{ig}\sigma_g + \nu^*_{ig}$, woraus $\nu^*_{ig} = \nu_{ig} - \lambda^*_{ig}\mu_g = \nu_{ig} - \mu_g\lambda_{ig}/\sigma_g$ folgt.

kleineren Abweichungen stärker bestraft werden, so führt dies dazu, dass die Modellparameter eher durch die kleineren als durch die größeren Abweichungen definiert werden. Damit führt das invariance alignment dazu, dass es sehr viele kleine Abweichungen der Itemladungen und (transformierten) Item Intercepts zwischen den Ländern gibt und nur wenige sehr große Abweichungen. Man setzt daher eine datengetriebene Methode ein, die partieller Invarianz entspricht. Daher findet diese Methode praktisch diejenigen Items, auf denen die Gruppenvergleiche basieren. Zusätzlich schlagen Asparouhov und Muthén (2014) Effektgrößen der praktischen Invarianz vor, die auf R^2-Maßen beruhen.

Im R-Paket sirt (Robitzsch, 2015) ist das invariance alignment in einer abgewandelten Form in der Funktion `invariance.alignment` implementiert. Zunächst wird anstelle der simultanen Schätzung von Mittelwerten und Standardabweichungen das Schätzproblem zweischrittig ausgeführt, in dem in einem ersten Schritt Standardabweichungen und in einem zweiten Schritt Mittelwerte geschätzt werden. Dadurch konnte in einigen Beispielen ein besseres Konvergenzverhalten erhalten werden. Außerdem ist in der Funktion `invariance.alignment` eine allgemeinere Zielfunktion f implementiert. Anstelle der Wahl von $f(x) = (x^2 + \varepsilon)^{1/4}$ in Asparouhov und Muthén (2014) setzen wir $f_{a,p} = ((x/a)^2 + \varepsilon)^p$ ein, was mit $a = 1$ und $p = 1/4$ zur originalen Definition führt. Durch die Potenz p kann man das Ausmaß der Robustheit der Zielfunktion definieren. Dabei führt $p = 1$ zur quadratischen Zielfunktion und damit zu einer ähnlichen Methode wie nach Haberman (2009). Durch die Wahl des Skalenparameters a kann man die Bedeutsamkeit von Effekten der Nichtinvarianz steuern. Postuliert man, dass Abweichungen in Ladungen von bereits .30 relevant sind, so könnte man $a = .30$ wählen und Abweichungen größer als .30 werden in der Zielfunktion mit $p = 1/4$ weniger stark bestraft, während diese stärker bei Werten kleiner als .30 bestraft werden. Gruppenvergleiche basieren dann tendenziell auf Items, die „mittlere“ Abweichungen in Ladungen mit Werten kleiner als .30 besitzen. Analog kann ein für die jeweilige Fragestellung bedeutsamer Skalenparameter a auch für Item Intercepts durch den Anwender vorgegeben werden.

Anstelle der Methode des invariance alignment schlagen Muthén und Asparouhov (2012) in einer Bayesianischen Ansatz informative Priorverteilungen zur Modellierung von Effekten der Nichtinvarianz (d.h. Unterschiede zwischen Itemladungen oder Item Intercepts zwischen Gruppen) vor (siehe auch Van De Schoot et al., 2013). Wie im Modell des invariance alignment dominieren bei der Anpassung der Faktormodelle nur „bedeutsame“ Effekte die Priorverteilung und werden „signifikant“ (siehe hierfür Muthén & Asparouhov, 2012). De Jong et al. (2007) oder Fox und Verhagen (2010) (siehe auch Fox, 2010) modellieren hierarchische Verteilungen für Itemparameter für mehrere Gruppen im Rahmen Bayesianischer IRT-Modelle. Dadurch können Gruppenunterschiede auch bei Existenz von Nichtinvarianz identifiziert werden. Für Itemladungen und Item Intercepts kann daher eine hierarchische bivariate Normalverteilung $(\lambda_{ig}, \nu_{ig}) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Psi}_1)$ angenommen werden (siehe Fox & Verhagen, 2010), wobei $\boldsymbol{\mu}_i$ die gemeinsamen Itemparameter (über Gruppen hinweg) für Item i bezeichnet. Für diese Itemparameter wird auf der zweiten Ebene wiederum eine hierarchische Verteilung $\boldsymbol{\mu}_i \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Psi}_2)$ angenommen, die die Verteilung aller Itemparameter modelliert. Bei einem festen Test (z.B. einer Skala mit fünf Items im PISA-Fragebogen) ist die zweite Stufe der Modellierung vielleicht weniger bedeutsam. Wenn die Passung der Items im Sinne der Invarianz zwischen den Gruppen variiert, so wird man vielleicht eher die Kovarianzmatrix $\boldsymbol{\Psi}_{1g}$ gruppenspezifisch schätzen. Will man

Items detektieren, die unterschiedlich zwischen Gruppen funktionieren, so würde man $\mathbf{\Psi}_{1i}$ itemspezifisch schätzen und Items mit großen Varianzen näher studieren.

In Ländervergleichsstudien wie in PISA wird Invarianz für Skalen praktisch niemals erfüllt sein. Daher schlagen Glas und Jehangir (2014) im Sinne einer *praktischen Invarianz* vor, die Stabilität von Aussagen über Ländern unter verschiedenen Skalierungsmodellen zu untersuchen (vgl. auch Glas, 2009). Eine Skala ist dabei praktisch invariant, wenn sich Länderreihenfolgen unter verschiedenen Skalierungsmodellen nicht ändern. Die Veränderung von Parametern in verschiedenen Modellen diskutiert Oberski (2014) im Konzept des *expected parameter change* (siehe auch Saris, Satorra & Van der Veld, 2009).

Längsschnittanalysen

Liegen Längsschnittdaten vor, so stellt sich ebenso die Frage der Bedeutung der Invarianz (Millsap, 2010). Wir haben unter einer Item Sampling (d.h. Domain Sampling) Annahme in Kapitel 4 argumentiert, dass wir Invarianz von Itemparametern *nicht* als notwendige Voraussetzung für die Messung längsschnittlicher Veränderung ansehen. Im Rahmen von Kompetenztests sehen wir einen vorliegenden Test als Itemstichprobe einer größeren Itemdomäne an. Berichtete Effektgrößen der Veränderung hängen daher auch bei unendlich großen Personenstichproben von der Itemauswahl ab.

Wir gehen davon aus, dass für Items zu zwei Zeitpunkten Itemschwierigkeiten b_{i1} und b_{i2} vorliegen. Dabei soll die Anpassung mit einem Rasch-Modell unter $E(\theta_{p1}) = E(\theta_{p2}) = 0$ erfolgen. Veränderung in Fähigkeiten wird dann indirekt in Veränderung in Itemschwierigkeiten abgebildet. Für jedes Item i kann man dann eine itemspezifische längsschnittliche Veränderung $d_i = b_{i1} - b_{i2}$ definieren. Die mittlere Veränderung ist dann $d = 1/I \cdot \sum_i d_i$. Formal entspricht dieses Modell der Annahme, dass unter Item Sampling $d_i \sim N(d, \sigma^2_{DIF})$ gilt. Dabei bezeichnet σ^2_{DIF} die längsschnittliche DIF-Varianz. Für I unabhängig „gezogene“ Items aus der Itemdomäne ist dann der durch Item Sampling bedingte Standardfehler von d durch $SE(d) = \sigma_{DIF}/\sqrt{I}$ gegeben (siehe Kapitel 4). Typischerweise wird die Unsicherheit der Erfassung längsschnittlicher Veränderung bei großen Itemstichproben kleiner. Ich erachte die Beurteilung der Invarianz mittels Modellfit-Statistiken oder Modellvergleichen für nicht sinnvoll, da diese keine Aussagen in der Metrik des interessierenden Parameters (z.B. Cohen's d für die Veränderung) liefern.

Gerade bei instruktionssensitiven Items und stärker curricular angelegten Tests (Polikoff, 2010) scheint die Annahme, dass die Items in einem homogenen Ausmaß Veränderung abbilden und daher Invarianz gilt, höchst unplausibel. Ein Test auf längsschnittlichen DIF liefert auch keine Aussage darüber, ob man mit einem Test „wahre Veränderung“ abbildet. Dies ist von der „repräsentativen“ Auswahl der Items abhängig und keine psychometrische Frage. Invarianztests liefern nur Aussagen über die Homogenität. Validitätsbelege für einen eingesetzten Test zur Erfassung von längsschnittlicher Veränderung könnte man durch Vergleich mit Effektgrößen publizierter Studien oder konfirmatorischen DIF-Analysen durchführen, mit denen a priori postuliert wird, für welche Items (bzw. Itemgruppen) besonders starke Änderungen in Itemschwierigkeiten erwartet werden (siehe De Boeck & Wilson, 2004 für entsprechende IRT-Modelle).

Auch im Fall einer festen Itemmenge würde ich Invarianz nicht als Voraussetzung für die Analyse längsschnittlicher Veränderung ansehen (siehe Kapitel 5 bzw. Robitzsch

et al., 2011). Die DIF-Effekte können konstruktinhärente Varianzanteile darstellen, die relevanten Sekundärdimensionen (Roussos & Stout, 1996) oder itemspezifischen Residuen entsprechen (McCrae, 2015). Wenn bei einem Test zur Messung der Mathematikkompetenz ein Item der schriftlichen Division in der fünften Klasse eine höhere Ladung als in der sechsten Klasse besitzt, so scheint dies curricular begründbar. Einzelne Facetten der Mathematikkompetenz (hier der schriftlichen Division) können im Hinblick auf die globale Kompetenz ihr „Gewicht" ändern (McCrae, 2015). Dies stellt kein Problem dar, wenn die Menge der Facetten, auf die sich die Erfassung der längsschnittlichen Veränderung bezieht, definiert ist.

Linkingfehler in Large Scale Assessments

Abschließend möchten wir die Konsequenzen des Domain Samplings für Large Scale Assessments (wie PISA oder PIRLS/TIMSS) diskutieren (basierend auf Robitzsch, 2011a und Robitzsch, 2012). Wir beziehen uns dabei primär auf die Bedeutung von Länderunterschieden. In Abbildung 7.3 ist ein typisches Design zweier PISA-Erhebungen der Jahre 2006 und 2009 illustriert. Dabei gibt es Items (die sog. *Linkitems*), die zu beiden Erhebungen eingesetzt werden und Items, die nur zu jeweils einem der beiden Zeitpunkten vorgelegt werden.

In PISA werden in jeder Studie internationale Itemparameter bestimmt, die dann für die Skalierung der Kompetenzwerte aller Länder eingesetzt werden (siehe OECD, 2014 und darin auch diskutierte Ausnahmen). Zu jedem der Zeitpunkte werden dann nationale Mittelwerte μ_1 und μ_2 für alle Teilnehmerländer g bestimmt. Von Interesse ist dabei primär die Unsicherheit *originaler Trendschätzungen* $\Delta_{orig} = \mu_2 - \mu_1$ (siehe Monseur & Berezner, 2007; Carstensen, Prenzel & Baumert, 2008). Durch Itemauswahl bedingte Unsicherheit (Michaelides & Haertel, 2004; Michaelides, 2010) wird mitunter auch als *Linkingfehler* bezeichnet (Monseur & Berezner, 2007; Monseur, Sibberns & Hastedt, 2008; siehe auch Robitzsch, 2010)

Originale Trendschätzungen in Large-Scale Assessments beruhen im Allgemeinen auf internationalen Itemparametern, die für alle Teilnehmerländer zugleich gelten. Typischerweise wird jedoch länderspezifischer DIF existieren, so dass „wahre" nationale Itemparameter von dem internationalen Itemparametern abweichen und diese Variabilität (bzw. dieser Modellfehler) Variabilität in Ländermittelwerten und für Trendschätzungen besitzt.

Für die kommenden Überlegungen nehmen wir an, dass wir für beide Studien ein Rasch-Modell anpassen, so dass wir als Itemparameter nur Itemschwierigkeiten verwenden. Dabei gehen wir davon aus, dass länderspezifische Itemparameter b_{it} zu den Zeitpunkten $t = 1$ und $t = 2$ existieren. Internationale Itemparameter bezeichnen wir mit β_{it}. Wir nehmen an, dass die Parameter b_{it} und β_{it} zentriert sind. Folgendes hierarchisches Modell wird dann angenommen

$$\begin{aligned} \beta_{it} &= \gamma_i + \epsilon_{it} \\ b_{it} &= u_i + e_{it} \end{aligned} \tag{7.85}$$

Wir spezifizieren eine bivariate Normalverteilungen für (u_i, γ_i) mit Kovarianzmatrix $\mathbf{\Sigma}_1$ und (e_{it}, ϵ_{it}) mit Kovarianzmatrix $\mathbf{\Sigma}_2$ [23]. Die Effekte $b_{it} - \beta_{it} = (u_i - \gamma_i) + (e_{it} - \varepsilon_{it})$

[23] Dabei nehmen wir an, dass die Verteilungen (e_{i1}, ϵ_{i1}) und (e_{i2}, ϵ_{i2}) übereinstimmen.

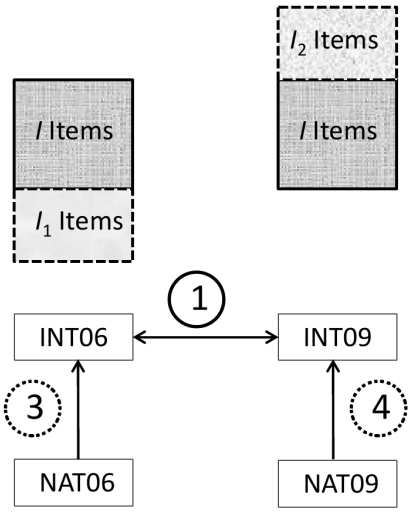

Abbildung 7.3: *Testdesign für eine originale Trendschätzung, wenn die Linkitems bei beiden Erhebungszeitpunkten echte Teilmengen der gesamten Itemmenge darstellen*

kennzeichnen einen länderspezifischen querschnittlichen DIF für Studie t. Ein über beide Studien hinweg auftretender querschnittlicher DIF wird durch $u_i - \gamma_i$ parametrisiert, eine Interaktion zwischen Ländern und Zeitpunkten durch die Differenz $e_{it} - \varepsilon_{it}$. Durch diesen Term wird abgebildet, dass ein bestimmtes Item beispielsweise zum ersten Zeitpunkt DIF aufweisen kann, zum zweiten Zeitpunkt jedoch nicht.

Ein internationaler *Item Drift* ist durch $\beta_{i2} - \beta_{i1} = \epsilon_{i2} - \epsilon_{i1}$ definiert. Die Differenz $b_{i2} - b_{i1} = e_{i2} - e_{i1}$ ist der länderspezifische Item Drift. Im Modell wird angenommen, dass die Items unabhängig voneinander sind (bzw. unabhängig „funktionieren“). Da PISA-Items häufig in Testlets auftreten, ist diese Annahme im Hinblick auf die Ableitung von Fehlern eher als konservativ zu bewerten.

Die Unsicherheit des Ländermittelwertes zum ersten Zeitpunkt (2006) kann gemäß (7.85) als querschnittliche DIF-Varianz ermittelt werden. Die durch Item Sampling der $I + I_1$ verwendeten Items verursachte Variabilität des Ländermittelwertes (als Mittelwert von $b_{i1} - \beta_{i1} = (u_i - \gamma_i) - (e_{i1} - \epsilon_{i1})$) für die erste Studie beträgt

$$[SE(\mu_1)]^2 = \frac{Var(u_i - \gamma_i) + Var(e_{it} - \epsilon_{it})}{I + I_1} = \frac{\sigma^2_{u-\gamma} + \sigma^2_{e-\epsilon}}{I + I_1} \tag{7.86}$$

Analog kann der Standardfehler für den zweiten Zeitpunkt abgeleitet werden.

Für originale Trendschätzungen $\mu_{g2} - \mu_{g1}$ wird empfohlen, den Linkingfehler bei Signifikanztestungen zusätzlich zum Person Sampling zu berücksichtigen (OECD, 2014; siehe auch Monseur & Berezner, 2007). Mit dem Modell (7.85) basiert der Linkingfehler auf der Varianz der internationalen Itemparameter, die durch $Var(\beta_{i2} - \beta_{i1}) = 2 \cdot Var(\epsilon_{it})$ gegeben ist. Demzufolge erhält man für den Linkingfehler auf Basis der Ankeritems

$$[\textit{Linkingfehler}]^2 = 2 \cdot \frac{\sigma^2_\epsilon}{I} \tag{7.87}$$

Ich argumentiere, dass (7.87) die Unsicherheit der originalen Trendschätzung nicht korrekt abbildet und leiten aus (7.85) eine Formel für den Standardfehler von Δ_{orig} ab.

Für die I Linkitems können wir dabei schreiben

$$(b_{i2} - \beta_{i2}) - (b_{i1} - \beta_{i1}) = (e_{i2} - \epsilon_{i2}) - (e_{i1} - \epsilon_{i1}) \tag{7.88}$$

Für den ersten Zeitpunkt ergibt sich jedoch für die I_1 Nicht-Linkitems die Differenz

$$(b_{i1} - \beta_{i1}) = (u_i - \gamma_i) - (e_{i1} - \epsilon_{i1}) \tag{7.89}$$

Anhand von (7.88) und (7.89) wird plausibel, dass für die Genauigkeit von Trendschätzungen sowohl die Varianzquelle der Interaktion aus Land und Zeitpunkt ($Var(e_{i2} - \epsilon_{i2})$) als auch ein über beide Zeitpunkte auftretender Länder-DIF ($Var(u_i - \gamma_i)$) bedeutsam ist. Dabei ergibt sich mit einfacher Algebra die Beziehung

$$[SE(\Delta_{orig})]^2 = \left(\frac{(w_2 - w_1)^2}{I} + \frac{1 - w_2}{I + I_2} + \frac{1 - w_1}{I + I_1} \right) \cdot \sigma^2_{u-\gamma} + \left(\frac{1}{I + I_1} + \frac{1}{I + I_2} \right) \cdot \sigma^2_{e-\epsilon} \tag{7.90}$$

Dabei haben wir als Anteile der Linkitems die Definitionen $w_1 := I/(I + I_1)$ sowie $w_2 := I/(I + I_2)$ verwendet. In diesen Standardfehler geht also die internationale bedingte Variabilität $\sigma^2_{u-\gamma}$ in Itemparametern und die national bedingte Variabilität σ^2_{e2-e1} ein. Typischerweise wird die in Large Scale Assessments verwendete Formel (7.87) für den Linkingfehler von unserer vorgeschlagenen Formel (7.90) verschieden sein. Spezialisieren wir nun (7.90) auf den Fall, dass die Items in beiden Studien identisch sind, so folgt mit $I_1 = I_2 = 0$ wegen $w_1 = w_2 = 1$ gerade

$$[SE(\Delta_{orig})]^2 = \left(\frac{2}{I} \right) \cdot \sigma^2_{e-\epsilon} \tag{7.91}$$

woraus der Unterschied zu (7.87) deutlich wird. Demzufolge sollte der international berichtete Linkingfehler (7.87) nicht für die Signifikanztestung originaler Trendschätzungen eingesetzt werden. Nur im Fall von $\sigma^2_{e-\epsilon} = \sigma^2_{\epsilon}$ gilt die Gleichheit der beiden Standardfehler. D.h. aber, dass $\sigma^2_e = 2 \cdot Cov(e, \epsilon)$ erfüllt sein muss. Im Fall von $\sigma^2_e = \sigma^2_{\epsilon}$ bedeutet dies aber, dass $Cor(e, \epsilon) = 1/2$ gelten muss.

Nehmen wir nun einen zweiten Spezialfall an, der dem PISA-Design für die Lesekompetenz in den Studien 2006 und 2009 entspricht. Alle Items aus dem Jahr 2006 sind auch 2009 eingesetzt worden, so dass $I_1 = 0$ folgt. Außerdem sind 2009 viele neu entwickelte Items mit $I_2 \gg I$ eingesetzt worden. Dann ist die Variabilität der originalen Trendschätzung gegeben durch

$$[SE(\Delta_{orig})]^2 = \left(\frac{(1 - w_2)^2}{I} + \frac{1 - w_2}{I + I_2} \right) \cdot \sigma^2_{u-\gamma} + \left(\frac{1}{I} + \frac{1}{I + I_2} \right) \cdot \sigma^2_{e-\epsilon} \tag{7.92}$$

Wenn die Anzahl I der Linkitems viel kleiner als die Anzahl I_2 der Nicht-Linkitems ist, so folgt (mit $I_2 \to \infty$)

$$[SE(\Delta_{orig})]^2 \approx \left(\frac{1}{I} \right) \cdot \sigma^2_{u-\gamma} + \left(\frac{1}{I} \right) \cdot \sigma^2_{e-\epsilon} \tag{7.93}$$

Man erkennt an (7.93), dass für Fehler in der originalen Trendschätzung auch querschnittlicher Länder-DIF $\sigma^2_{u-\gamma}$ eine Rolle spielen kann, wenn man wenige Linkitems verwendet.

Verwendet man dieselben Items in beiden Studien, so verschwindet diese Fehlerquelle (siehe (7.90)).

Alternativ zu originalen Trendschätzungen werden auch *marginale Trendschätzungen* Δ_{marg} (Carstensen et al., 2008) diskutiert, die einen länderspezifischen Trend ausschließlich auf Basis länderspezifischer Itemparameter berechnen. Bei dieser Trendschätzung wird jedes Land praktisch an sich selbst verankert, so dass der „Fehler" bei der Verwendung internationaler Itemparameter entfällt. Praktisch wird damit die Veränderung eines Landes auf Basis der Linkitems der beiden Erhebungen bestimmt. Unsicherheit in der Mittelwertdifferenz $\mu_{i2} - \mu_{i1}$ aufgrund der Itemauswahl sind dabei nur auf die Differenzen $b_{i2} - b_{i1} = e_{i2} - e_{i1}$ zurückzuführen. Damit ergibt sich eine Schätzung der Variabilität der marginalen Trendschätzung als

$$[SE(\Delta_{marg})]^2 = \left(\frac{2}{I}\right) \cdot \sigma_e^2 \tag{7.94}$$

Die marginale Trendschätzung kann verwendet werden, um die Robustheit und Validität der in den offiziellen Berichten verwendeten originalen Trendschätzung zu untersuchen (Robitzsch, 2012).

Aus einer Perspektive der Minimierung von Standardfehlern (aufgrund von Item Sampling) kann die marginale Trendschätzung effizienter als die originale Trendschätzung sein. Longford (2003, 2008) leitet einen sog. *composite estimator* als Konvexkombination von vorliegenden als unverzerrt angenommenen Schätzern ab, der eine minimale Varianz (d.h. einen minimalen RMSE) besitzt. Im Falle der Trendschätzung sind in der unendlichen Itempopulation sowohl die originale als auch die marginale Trendschätzung unverzerrt, so dass man einen composite estimator der Form $\Delta_{comp} = w \cdot \Delta_{orig} + (1 - w) \cdot \Delta_{marg}$ mit einem Gewicht w ($0 \leq w \leq 1$) bestimmt.

Die Bedeutung der Itemauswahl im Large Scale Assessment soll an einer Studie illustriert werden. Für Österreich wurde in der PISA-Studie ein Leistungsabfall (originale Trendschätzung) der Lesekompetenz von PISA 2006 nach PISA 2009 von 20 Punkten berichtet (Schwantner & Schreiner, 2010). Dabei zeigten Analysen des Autors, dass der Leistungsabfall auf den Linkitems weniger dramatisch ausfiel. Tatsächlich zeigt die ermittelte marginale Trendschätzung nur einen nichtsignifikanten Leistungsabfall von 4 Punkten. Eine Anpassung des hierarchischen Modells für Itemparameter (7.85) wurde genutzt, um einen composite estimator abzuleiten. Dabei ergab sich ein Gewicht von $w = .44$ für die originale Trendschätzung, so dass insgesamt der „effizienteste" composite estimator Δ_{comp} ein Leistungsabfall von 11 Punkten ($.44 \cdot 20 + (1 - .44) \cdot 4 = 11$) ermittelt wurde.

Die Befunde dieses Abschnittes besitzen auch Bedeutung für längsschnittliche Studien mit Linkitems sowie nur zu einzelnen Zeitpunkten eingesetzten Items. Wenn die mittlere Veränderung zweier Gruppen (z.B. Schulformen) ermittelt werden soll, so wird dann längsschnittlicher Item Drift als auch querschnittlicher DIF bei Itemauswahl Variabilität in Effektgrößen verursachen. Bei wenigen Linkitems kann diese durch Item Sampling verursachte Variabilität mitunter bedeutsamer als die Variabilität durch Person Sampling sein (vgl. auch Kapitel 4).

Literaturverzeichnis

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation*, *31*, 162–172.

Adams, R. J., Wilson, M. & Wang, W.-C. (1997). The multidimensional random coefficient multinomial logit model. *Applied Psychological Measurement*, *21*, 1–23.

Adams, R. J. & Wu, M. L. (2007). The mixed-coefficients multinomial logit model: A generalized form of the Rasch model. In M. von Davier & C. Carstensen (Hrsg.), *Multivariate and mixture distribution Rasch models* (S. 57–75). New York: Springer.

Afflerbach, P. D., P. & Paris, S. (2008). Clarifying differences between reading skills and reading strategies. *The Reading Teacher*, *61*, 364–373.

Agarwal, D., Zhang, L. & Mazumder, R. (2011). Modeling item-item similarities for personalized recommendations on Yahoo! front page. *Annals of Applied Statistics*, *5*, 1839–1875.

Aitkin, M. & Aitkin, I. (2011). *Statistical modeling of the national assessment of educational progress.* New York: Springer.

Albano, A. D. (2013). Multilevel modeling of item position effects. *Journal of Educational Measurement*, *50*(4), 408–426.

Alexandrowicz, R. & Matschinger, H. (2008). Estimation of item location effects by means of the generalized logistic regression model: A simulation study and an application. *Psychology Science*, *50*(1), 64–74.

Algina, J., Keselman, H. J. & Penfield, R. D. (2005). Effect sizes and their intervals: The two-level repeated measures case. *Educational and Psychological Measurement*, *65*, 241–258.

Alisch, L.-M. (2002). Durchschnitte - Ergebnisse der Ergodentheorie zu ihrer Bedeutung für die empirische Forschung. *Empirische Pädagogik*, *16*, 79-94.

Alonso, A., Litière, S. & Laenen, A. (2010). A note on the indeterminacy of the random-effects distribution in hierarchical models. *The American Statistician*, *64*, 318–324.

Anderson, C. J., Li, J. & Vermunt, J. K. (2007). Estimation of models in a Rasch family for polytomous items and multiple latent variables. *Journal of Statistical Software*, *20*(6), 1–36.

Anderson, C. J. & Vermunt, J. K. (2000). Log-multiplicative association models as latent variable models for nominal and/or ordinal data. *Sociological Methodology*, *30*(1), 81–121.

Anderson, C. J. & Yu, H.-T. (2007). Log-multiplicative association models as item response models. *Psychometrika*, *72*(1), 5–23.

Anderson, J. C. & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*(3), 411–423.

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis.* New York: Wiley.

Angrist, J. D. & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion.* Princeton University Press.

Arellano, M. & Bonhomme, S. (2011). Nonlinear panel data analysis. *Annual Review of Economics*, *3*, 395–424.

Arellano, M. & Hahn, J. (2006). *A likelihood-based approximate solution to the incidental parameter problem in dynamic nonlinear models with multiple effects* (Tech. Rep.).

Artelt, C. & Dörfler, T. (2010). Förderung von Lesekompetenz als Aufgabe aller Fächer. Forschungsergebnisse und Anregungen für die Praxis. In Bayerisches Staatsministerium für Unterricht und Kultus (Hrsg.), *ProLesen. Auf dem Weg zur Leseschule* (S. 13–36). Donauwörth: Auer.

Asparouhov, T. & Muthen, B. (2008). Multilevel mixture models. In G. R. Hancock & K. M. Samuelsen (Hrsg.), *Advances in latent variable mixture models* (S. 27–51). Information Age Charlotte, NC.

Asparouhov, T. & Muthén, B. (2010). *Plausible values for latent variables using Mplus* (Tech. Rep.). Mplus Technical Report. http://www. statmodel.com/download/Plausible.pdf.

Asparouhov, T. & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling*, *21*(4), 495–508.

Baayen, R. H., Davidson, D. J. & Bates, D. M. (2008). Mixed effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Babcock, B. (2011). Estimating a noncompensatory IRT model using metropolis within Gibbs sampling. *Applied Psychological Measurement*, *35*(4), 317–329.

Bacci, S. & Bartolucci, F. (2013). *A multidimensional latent class IRT model for non-ignorable missing responses* (Tech. Rep.). http://ssrn.com/abstract=2213721.

Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transactions*, *22*, 1145–1146.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, *71*, 135–171.

Bai, J. & Li, K. (2012a). *Maximum likelihood estimation and inference for approximate factor models of high dimension* (Tech. Rep.).

Bai, J. & Li, K. (2012b). Statistical analysis of factor models of high dimension. *Annals of Statistics*, *40*, 436–465.

Bai, J. & Liao, Y. (2013). *Efficient estimation of approximate factor models via regularized maximum likelihood* (Tech. Rep.).

Bai, J. & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, *70*, 191–221.

Bai, J. & Ng, S. (2008). Large dimensional factor analysis. *Foundations and Trends in Economics*, *3*, 89–163.

Bai, Z. & Silverstein, J. W. (2010). *Spectral analysis of large dimensional random matrices.* New York: Springer.

Ballou, D. (2009). Test scaling and value-added measurement. *Education Finance and Policy*, *4*, 351–383.

Bartolucci, F. (2007). A class of multidimensional IRT models for testing unidimensionality and clustering items. *Psychometrika*, *72*(2), 141–157.

Bartolucci, F., Montanari, G. & Pandolfi, S. (2012). Dimensionality of the latent structure and item selection via latent class multidimensional IRT models. *Psychometrika*, *77*(4), 782–802.

Bazán, J. L., Branco, M. D. & Bolfarine, H. (2006). A skew item response model. *Bayesian Analysis*, *1*(4), 861–892.

Beaujean, A. A. (2014). *Latent variable modeling using R.* New York: Routledge.

Becker, M., Lüdtke, O., Trautwein, U. & Baumert, J. (2006). Leistungszuwachs in Mathematik: Evidenz für einen Schereneffekt im mehrgliedrigen Schulsystem? *Zeitschrift für Pädagogische Psychologie*, *20*, 233–242.

Berk, R., Brown, L., Buja, A., George, E., Pitkin, E., Zhang, K. & Zhao, L. (2014). Misspecified mean function regression making good use of regression models that are wrong. *Sociological Methods & Research*, *43*(3), 422–451.

Berk, R., Brown, L. & Zhao, L. (2010). Statistical inference after model selection. *Journal of Quantitative Crimonology*, *26*, 217–236.

Bertoli-Barsotti, L., Lando, T. & Punzo, A. (2014). Estimating a Rasch model via fuzzy empirical probability functions. In D. Vicari, A. Okada, G. Ragozini & C. Weihs (Hrsg.), *Analysis and modeling of complex data in behavioral and social sciences* (S. 29–36). New York: Springer.

Bertoli-Barsotti, L. & Punzo, A. (2013). Rasch analysis for binary data with nonignorable nonresponses. *Psicológica*, *34*(1), 97–123.

Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion.* München: Pearson.

Bickel, P. J. & Li, B. (2006). Regularization in statistics (with discussion). *Test*, *15*(2), 271–344.

Binder, D. A. & Roberts, G. (2012). Design- and model-based inference for model parameters. In C. R. Rao (Hrsg.), *Handbook of Statistics, Volume 29B: Inference and Analysis* (S. 33–54). North Holland: Elsevier.

Bingham, E., Kabán, A. & Fortelius, M. (2009). The aspect Bernoulli model: Multiple causes of presences and absences. *Pattern Analysis and Applications*, *12*(1), 55–78.

Blei, D. M. & Lafferty, J. D. (2007). A correlated topic model of science. *Annals of Applied Statistics*, *1*(1), 17–35.

Blömeke, S., Gustafsson, J.-E. & Shavelson, R. J. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie*, *223*(1), 3–13.

Bock, R. D., Brennan, R. L. & Muraki, E. (2002). The information in multiple ratings. *Applied Psychological Measurement*, *26*, 364–375.

Bock, R. D. & Moustaki, I. (2007). Item response theory in a general framework. In C. R. Rao & S. Sinharay (Hrsg.), *Handbook of Statistics, Volume 26: Psychometrics* (S. 469–513). North Holland: Elsevier.

Böhme, K. & Bremerich-Vos, A. (2009). Diagnostik der Rechtschreibkompetenz in der Grundschule – Konstruktprüfung mittels Fehler- und Dimensionsanalysen. In A. Bremerich-Vos, D. Granzer & O. Köller (Hrsg.), *Bildungsstandards Deutsch und Mathematik* (S. 340–365). Weinheim: Beltz.

Böhme, K. & Robitzsch, A. (2009). Methodische Aspekte der Erfassung der Lesekompetenz. In A. Bremerich-Vos, D. Granzer & O. Köller (Hrsg.), *Bildungsstandards Deutsch und Mathematik* (S. 250–289). Weinheim: Beltz.

Bolt, D. M. & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement*, *27*, 395-414.

Bolt, D. M., Wollack, J. A. & Suh, Y. (2012). Application of a multidimensional nested logit model to multiple-choice test items. *Psychometrika*, *77*(2), 339–357.

Bond, T. G. & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah: Lawrence Erlbaum.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics.* Cambridge: Cambridge University Press.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*, 425–440.

Borsboom, D. & Mellenbergh, G. J. (2007). Test validity in cognitive assessment. In J. Leighton & M. Gierl (Hrsg.), *Cognitive diagnosis assessment for education: Theory and applications* (S. 85–116). Cambridge: Cambridge University Press.

Borsboom, D. & Zand Scholten, A. (2008). The Rasch model and conjoint measurement theory from the perspective of psychometrics. *Theory & Psychology*, *18*(1), 111–117.

Bouchaud, J.-P. & Potters, M. (2009). Financial applications of random matrix theory: a short review. *arXiv preprint arXiv:0910.1205.*

Bradlow, E. T., Wainer, H. & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168.

Braeken, J. (2011). A boundary mixture approach to violations of conditional independence. *Psychometrika*, *76*, 57–65.

Braeken, J., Kuppens, P., De Boeck, P. & Tuerlinckx, F. (2013). Contextualized personality questionnaires: A case for copulas in structural equation models for categorical data. *Multivariate Behavioral Research*, *48*(6), 845–870.

Braeken, J., Tuerlinckx, F. & De Boeck, P. (2007). Copula functions for residual dependency. *Psychometrika*, *72*, 393–411.

Brandt, S. (2008). Estimation of a Rasch model including subdimensions. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *1*, 51–69.

Brandt, S. (2010). Estimating tests using subtests. *Journal of Applied Measurement*, *11*, 352–367.

Brandt, S. & Duckor, B. (2013). Increasing unidimensional measurement precision using a multidimensional item response model approach. *Psychological Test and Assessment Modeling*, *55*, 148–161.

Brechmann, E. C. & Joe, H. (2014). Parsimonious parameterization of correlation matrices using truncated vines and factor analysis. *Computational Statistics & Data Analysis*, *77*, 233–251.

Brennan, R. L. (2001a). *Generalizability theory*. New York: Springer.

Brennan, R. L. (2001b). Some problems, pitfalls, and paradoxes in educational measurement. *Educational Measurement: Issues and Practice*, *20*, 6–18.

Brennan, R. L. (2001c). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, *38*, 295–317.

Brennan, R. L. (2004). *Some perspectives on inconsistencies among measurement models* (CASMA Research Report Number 8). University of Iowa: Center for Advanced Studies in Measurement and Assessment (CASMA).

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Hrsg.), *Educational measurement* (S. 1–16). Westport: Praeger Publisher.

Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education*, *24*, 1–21.

Briggs, D. C. (2011). Cause or effect? Validating the use of tests for high-stakes inferences in education. In *Looking back. proceedings of a conference in honor of Paul W. Holland* (S. 131–147). New York: Springer.

Briggs, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement*, *50*(2), 204–226.

Briggs, D. C. & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement*, *28*, 3–14.

Briggs, D. C. & Wilson, M. (2007). Generalizability in item response modeling. *Journal of Educational Measurement*, *44*, 131–155.

Brock, W. A., Durlauf, S. N. & West, K. D. (2003). Policy evaluation in uncertain economic environments. *Brooking Papers on Economic Activity*, *1*, 235–301.

Brock, W. A., Durlauf, S. N. & West, K. D. (2007). Model uncertainty and policy evaluation: Some theory and empirics. *Journal of Econometrics*, *136*, 629–664.

Brunner, M., Nagy, G. & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality*, *80*, 796–846.

Brynjarsdóttir, J. & O'Hagan, A. (2014). Learning about physical parameters: The importance of model discrepancy. *Inverse Problems*, *30*(11).

Buckland, S. T., Burnham, K. P. & Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics*, *53*, 603–618.

Bühlmann, P. & Van De Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications.* New York: Springer.

Burnham, K. P. & Anderson, D. R. (2002). *Model selection and multimodel inference.* New York: Springer.

Burnham, K. P. & Anderson, D. R. (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*, 261–304.

Cameron, A. C. & Trivedi, P. K. (2005). *Microeconometrics.* New York: Cambridge University Press.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Hrsg.), *Educational measurement* (S. 221–256). Westport: Praeger Publisher.

Cao, J. & Stokes, S. L. (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika*, *73*(2), 209–230.

Carroll, R. J., Ruppert, D., Stefanski, L. A. & Crainiceanu, C. M. (2012). *Measurement error in nonlinear models: A modern perspective.* CRC Press.

Carstensen, C. H., Prenzel, M. & Baumert, J. (2008). Trendanalysen in PISA: Wie haben sich die Kompetenzen in Deutschland zwischen PISA 2000 und PISA 2006 entwickelt? *Zeitschrift für Erziehungswissenschaften*, *10*, 11–34.

Chaimongkol, S. (2005). *Modeling differential item functioning (DIF) using multilevel logistic regression models: A Bayesian perspective.* The Florida State University, PhD Thesis.

Chaimongkol, S., Huffer, F. & Kamata, A. (2007). An explanatory differential item functioning (DIF) model by the WinBUGS 1.4. *Songlanakarin Journal of Science and Technology*, *29*, 449–458.

Chamberlain, G. & Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, *51*, 1281–1304.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference (with discussion). *Journal of the Royal Statistical Society, Series A*, *158*, 419–466.

Cho, E. & Kim, S. (2015). Cronbach's coefficient alpha well known but poorly understood. *Organizational Research Methods*, *18*, 207–230.

Choppin, B. (1982). A fully conditional estimation procedure for Rasch model parameters. *Evaluation in Education*, *9*, 29–42.

Claeskens, G. & Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association*, *98*, 900–916.

Clarke, B. S. & Junker, B. W. (1991). *Inference from the product of marginals from a dependent likelihood* (Tech. Rep.). Department of Statistics, Carnegie Mellon University.

Clemen, R. T. & Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis*, *19*, 187–203.

Cliff, N. & Donoghue, J. R. (1992). Ordinal test fidelity estimated by an item sampling model. *Psychometrika*, *57*, 217–236.

Cliff, N. & Keats, J. A. (2003). *Ordinal measurement in the behavioral sciences.* Mahwah, NJ: Lawrence Erlbaum.

Clyde, M. & George, E. I. (2004). Model uncertainty. *Statistical Science*, *19*, 81–94.

Cole, D. A., Ciesla, J. A. & Steiger, J. H. (2007). The insidious effects of failing to include design-driven correlated residuals in latent-variable covariance structure analysis. *Psychological Methods*, *12*, 381–398.

Conijn, J. M., Emons, W. H., van Assen, M. A. & Sijtsma, K. (2011). On the usefulness of a multilevel logistic regression approach to person-fit analysis. *Multivariate Behavioral Research*, *46*(2), 365–388.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98.

Cramer, A. O., Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., Kendler, K. S. & Borsboom, D. (2012). Dimensions of normal personality as networks in search of equilibrium: You can't like parties if you don't like people. *European Journal of Personality*, *26*(4), 414–431.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* Wiley.

Cronbach, L. R. & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, *64*, 391-418.

Cudeck, R. & Henly, S. J. (1991). Model selection in covariance structures analysis and the problem of sample size: A clarification. *Psychological Bulletin*, *109*, 512–519.

Cunha, F., Heckman, J. J. & Schennach, S. M. (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, *78*, 883–931.

Davis, J. & Ferdous, A. (2005). *Using item difficulty and item position to measure test fatigue.* American Institutes for Research, Washington. Paris.

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*, 533–559.

De Boeck, P., Cho, S.-J. & Wilson, M. (2011). Explanatory secondary dimension modeling of latent differential item functioning. *Applied Psychological Measurement*, *35*(8), 583–603.

De Boeck, P. & Wilson, M. (Hrsg.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* New York: Springer.

de Gruijter, D. N. M. (1986). Small N does not always justify the Rasch model. *Applied Psychological Measurement*, *10*, 187–194.

de Jong, M. G., Steenkamp, J.-B. E. M. & Fox, J.-P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *Journal of Consumer Research*, *34*, 260–278.

de Leeuw, J. (1988). Models and techniques. *Statistica Neerlandica*, *42*, 91–98.

de Leeuw, J. & Verhelst, N. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational and Behavioral Statistics*, *11*, 183–196.

Debeer, D., Buchholz, J., Hartig, J. & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *Journal of Educational and Behavioral Statistics*, *39*(6), 502–523.

Debeer, D. & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement*, *50*, 164–185.

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement*, *36*(6), 447–468.

DeCarlo, L. T., Kim, Y. & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement*, *48*, 333–356.

Delaigle, A. & Meister, A. (2008). Density estimation with heteroscedastic error. *Bernoulli*, *14*(2), 562–579.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement*, *43*, 145–168.

Denoeux, T. (2011). Maximum likelihood estimation from fuzzy data using the EM algorithm. *Fuzzy Sets and Systems*, *183*, 72–91.

Denoeux, T. (2013). Maximum likelihood estimation from uncertain data in the belief function framework. *IEEE Transactions on Knowledge and Data Engineering*, *25*, 119–130.

Diamantopoulos, A. (2006). The error term in formative measurement models: Interpretation and modeling implications. *Journal of Modelling in Management*, *1*, 7–17.

Diamantopoulos, A., Riefler, P. & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research*, *61*, 1203–1218.

Domingue, B. (2014). Evaluating the equal-interval hypothesis with test score scales. *Psychometrika*, *79*(1), 1–19.

Doran, H., Bates, D., Bliese, P. & Dowling, M. (2007). Estimating the multilevel Rasch model: With the lme4 package. *Journal of Statistical Software*, *20(2)*.

Douglas, J. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika*, *62*, 7–28.

Douglas, J. (2001). Asymptotic identifiability of item response models. *Psychometrika*, *66*, 531–540.

Douglas, J. & Cohen, A. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement*, *25*(3), 234–243.

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society, Series B*, *57*, 45–97.

Draper, D. (2007). Bayesian multilevel analysis and MCMC. In J. de Leeuw & E. Meijer (Hrsg.), *Handbook of multilevel analysis* (S. 77–140). New York: Springer.

Draper, D., Hodges, J. S., Leamer, E. E., Morris, C. N. & Rubin, D. B. (1987). *A research agenda for assessment and propagation of model uncertainty* (Tech. Rep.). RAND Corporation Note Series.

Duke, N. K. & Carlisle, J. (2011). The development of comprehension. In M. L. Kamil, P. D. Pearson, E. B. Moje & P. P. Afflerbach (Hrsg.), *Handbook of Reading Research (Volume IV)* (S. 199–228). New York: Routledge.

Dunson, D. B. & Xing, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, *104*(487), 1042–1051.

Eckes, T. (2014). Examining testlet effects in the TestDaF listening section: A testlet response theory modeling approach. *Language Testing*, *31*, 39–61.

Eckes, T. (2015a). Lokale Abhängigkeit von Items im TestDaF-Leseverstehen. *Diagnostica*, *61*, 93–106.

Eckes, T. (2015b). Eine Replik auf den Kommentar zum Beitrag „Lokale Abhängigkeitt von Items im TestDaF-Leseverstehen“. *Diagnostica*, *61*, 110–111.

Edelman, A. & Rao, N. R. (2005). Random matrix theory. *Acta Numerica*, *14*, 233–297.

Edwards, J. R. & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*, 155–174.

Eid, M. & Koch, T. (2014). The meaning of higher-order factors in reflective-measurement models. *Measurement: Interdisciplinary Research & Perspectives*, *12*(3), 96–101.

Ellis, J. L. & Junker, B. W. (1997). Tail-measurability in monotone latent variable models. *Psychometrika*, *62*, 495–523.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*, 495–515.

Embretson, S. E. & Poggio, J. (2012). The impact of scaling and measurement methods on individual differences in growth. In B. Laursen, T. D. Little & N. A. Card (Hrsg.), *Handbook of developmental research methods* (S. 82–108). New York: Guilfort Press.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists.* New York: Taylor & Francis Group.

Emons, W. H., Sijtsma, K. & Meijer, R. R. (2005). Global, local, and graphical person-fit analysis using person-response functions. *Psychological Methods*, *10*(1), 101–119.

Epskamp, S., Maris, G., Waldorp, L. J. & Borsboom, D. (2015). Network psychometrics. In P. Irwing, D. Hughes & T. Booth (Hrsg.), *Handbook of psychometrics.* New York: Wiley.

Erosheva, E., Fienberg, S. & Lafferty, J. (2004). Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, *101*(Suppl. 1), 5220–5227.

Erosheva, E. A. (2005). Comparing latent structures of the grade of membership, Rasch, and latent class models. *Psychometrika*, *70*(4), 619–628.

Erosheva, E. A. (2006). *Latent class representation of the grade of membership model* (Technical Report No. 492). Seattle, University of Washington.

Erosheva, E. A., Fienberg, S. E. & Joutard, C. (2007). Describing disability through individual-level mixture models for multivariate binary data. *Annals of Applied Statistics*, *1*, 502–537.

Erosheva, E. A., Fienberg, S. E. & Junker, B. W. (2002). Alternative statistical models and representations for large sparse multi-dimensional contingency tables. *Annales de la Faculté des Sciences de Toulouse*, *11*(4), 485–505.

Fan, J., Fan, Y. & Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, *147*(1), 186–197.

Fan, J., Liao, Y. & Liu, H. (2015). Estimating large covariance and precision matrices. *arXiv preprint arXiv:1504.02995.*

Fan, J., Liao, Y. & Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society, Series B*, *75*(4), 603–680.

Feldt, L. S. (1997). Can validity rise when reliability declines? *Applied Measurement in Education*, *10*, 377–387.

Fernández-Val, I. (2009). Fixed effects estimation of structural parameters and marginal effects in panel probit models. *Journal of Econometrics*, *150*(1), 71–85.

Fernández-Val, I. & Vella, F. (2011). Bias corrections for two-step fixed effects panel data estimators. *Journal of Econometrics*, *163*(2), 144–162.

Ferrando, P. J. (2007). A Pearson-Type-VII item response model for assessing person fluctuation. *Psychometrika*, *72*(1), 25–41.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359-374.

Fischer, G. H. (1995a). Derivations of the Rasch model. In G. H. Fischer & I. W. Molenaar (Hrsg.), *Rasch models. foundations, recent developments, and applications* (S. 15-31). New York: Springer.

Fischer, G. H. (1995b). Linear logistic models for change. In G. H. Fischer & I. W. Molenaar (Hrsg.), *Rasch models. foundations, recent developments, and applications* (S. 157-180). New York: Springer.

Fischer, G. H. (2007). Rasch models. In C. R. Rao & S. Sinharay (Hrsg.), *Handbook of Statistics, Volume 26: Psychometrics* (S. 515–585). North Holland: Elsevier.

Fischer, G. H. & Molenaar, I. W. (Hrsg.). (1995). *Rasch models: Foundations, recent developments, and applications.* New York: Springer.

Formann, A. K. (1984). *Die Latent-Class-Analyse.* Weinheim: Beltz.

Formann, A. K. (1985). Constrained latent class models: Theory and applications. *British Journal of Mathematical and Statistical Psychology*, *38*, 87–111.

Formann, A. K. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association*, *87*, 476–486.

Formann, A. K. (1993). Some simple latent class models for attitudinal scaling in the presence of polytomous items. *Methodika*, *7*, 62–78.

Formann, A. K. (2007). (Almost) Equivalence between conditional and mixture maximum likelihood estimates for some models of the Rasch type. In M. von Davier & C. H. Carstensen (Hrsg.), *Multivariate and mixture distribution Rasch models* (S. 177–189). New York: Springer.

Formann, A. K. & Kohlmann, T. (1998). Structural latent class models. *Sociological Methods & Research*, *26*, 530–565.

Formann, A. K. & Kohlmann, T. (2002). Three-parameter linear logistic latent class analysis. In J. A. Hagenaars & A. L. McCutcheon (Hrsg.), *Applied latent class analysis* (S. 183–210). New York: Cambridge University Press.

Foster, E. M. & Kalil, A. (2008). New methods for new questions: obstacles and opportunities. *Developmental Psychology*, *44*, 301–304.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods.* Thousand Oaks: Sage.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications.* New York: Springer.

Fox, J.-P. & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*, 269–286.

Fox, J.-P. & Verhagen, A. J. (2010). Random item effects modeling for cross-national survey data. In E. Davidov, P. & Schmidt & J. Billiet (Hrsg.), *Cross-cultural analysis: Methods and applications* (S. 461–482). London: Routeledge Academic.

Freedman, D. A. et al.. (2008). Randomization does not justify logistic regression. *Statistical Science*, *23*(2), 237–249.

Frey, A., Hartig, J. & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice*, *28*(3), 39–53.

Frigerio, A., Giordani, A. & Mari, L. (2010). Outline of a general model of measurement. *Synthese*, *175*(2), 123–149.

Galyart, A. (2015). Interpreting mixed membership models: Implications of Erosheva's representation theorem. In E. M. Airoldi, D. Blei, E. A. Erosheva & S. E. Fienberg (Hrsg.), *Handbook of mixed membership models and its applications* (S. 39–65). Boca Raton: Chapman & Hall.

Garner, M. (2002). An eigenvector method for estimating item parameters of the dichotomous and polytomous Rasch models. *Journal of Applied Measurement*, *3*(2), 107–128.

Garthwaite, P. H. & Mubwandarikwa, E. (2010). Selection of weights for weighted model averaging. *Australian & New Zealand Journal of Statistics*, *52*, 362–382.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2004). *Bayesian data analysis.* Boca Raton, FL: Chapman & Hall.

Gelman, A. & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge: Cambridge University Press.

Gershunskaya, J., Jiang, J. & Lahiri, P. (2012). Resampling methods in surveys. In C. R. Rao (Hrsg.), *Handbook of Statistics, Volume 29B: Inference and Analysis* (S. 121–151). North Holland: Elsevier.

Ghahramani, Z., Mohamed, S. & Heller, K. (2015). A simple and general exponential family framework for partial membership and factor analysis. In E. M. Airoldi, D. Blei, E. A. Erosheva & S. E. Fienberg (Hrsg.), *Handbook of mixed membership models and its applications* (S. 67–88). Boca Raton: Chapman & Hall.

Gibbons, R. D. & Hedeker, D. R. (1992). Full-information bi-factor analysis. *Psychometrika*, *57*, 423–436.

Gifi, A. (1990). *Nonlinear multivariate analysis.* New York: Wiley.

Gignac, G. E. (2014). On the inappropriateness of using items to calculate total scale score reliability via coefficient alpha for multidimensional scales. *European Journal of Psychological Assessment*, *30*, 130–139.

Glas, C. A. W. (2009). *Core B report on the PISA 2009 background questionnaires.* EDU/PISA/GB(2009)9.

Glas, C. A. W. (2012a). *Estimating and testing the extended testlet model* (Tech. Rep.).

Glas, C. A. W. (2012b). Generalizability theory and item response theory. In T. Eggen & B. P. Veldkamp (Hrsg.), *Psychometrics in practice at RCEC* (S. 1–13).

Glas, C. A. W. & Jehangir, K. (2014). Modeling country-specific differential item functioning. In L. Rutkowski, M. von Davier & D. Rutkowski (Hrsg.), *Handbook of international large-scale assessment* (S. 97–115). Boca Raton: CRC Press.

Glas, C. A. W. & Pimentel, J. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement*, *68*, 907–922.

Gölitz, D., Roick, T. & Hasselhorn, M. (2006). *DEMAT 4: Deutscher Mathematiktest für vierte Klassen.* Göttingen: Hogrefe.

Goldstein, H. (1979). Consequences of using the Rasch model for educational assessment. *British Educational Research Journal*, 211–220.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology*, *33*, 234–246.

Goldstein, H. (2004). International comparisons of student attainment: Some issues arising from the PISA study. *Assessment in Education*, *11*, 319–330.

Goldstein, H. (2011). *Multilevel statistical models.* Chichester: Wiley.

Goldstein, H., Bonnet, G. & Rocher, T. (2007). Multilevel structural equation models for the analysis of comparative data on educational performance. *Journal of Educational and Behavioral Statistics*, *32*, 252–286.

Goldstein, H., Browne, W. & Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics*, *1*(4), 223–231.

Goldstein, H. & Wood, R. (1989). Five decades of item response modeling. *British Journal of Mathematical and Statistical Society*, *42*, 139–167.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576.

Graubard, B. I. & Korn, E. L. (1994). Regression analysis with clustered data. *Statistics in Medicine*, *13*, 509–522.

Green, S. B. & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika*, *74*, 155–167.

Greene, W. H. (2003). *Econometric analysis.* Pearson Education India.

Grömping, U. (1996). A note on fitting a marginal model to mixed effects log-linear regression data via GEE. *Biometrics*, 280–285.

Gruhl, J. & Erosheva, E. A. (2015). A tale of two (types of) memberships: Comparing mixed and partial membership with a continuous data example. In E. M. Airoldi, D. Blei, E. A. Erosheva & S. E. Fienberg (Hrsg.), *Handbook of mixed membership models and its applications* (S. 15–38). Boca Raton: Chapman & Hall.

Haberman, S. J. (2005). *Identifiability of parameters in item response models with unconstrained ability distributions* (ETS Research Report RR05-24). Princeton: ETS.

Haberman, S. J. (2007). *The information a test provides on an ability parameter* (ETS Research Report RR07-18). Princeton: ETS.

Haberman, S. J. (2009). *Linking parameter estimates derived from an item response model through separate calibrations* (ETS Research Report RR09-40). Princeton: ETS.

Haberman, S. J., Lee, Y. & Qian, J. (2009). *Jackknifing techniques for evaluation of equating accuracy* (ETS Research Report RR09-39). Princeton: ETS.

Haberman, S. J. & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika*, *75*, 209–227.

Haberman, S. J., von Davier, M. & Lee, Y. (2008). *Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous ability distributions* (ETS Research Report RR08-45). Princeton: ETS.

Habing, B. & Roussos, L. A. (2003). On the need of negative local dependence. *Psychometrika*, *68*, 435–451.

Hahn, J. & Newey, W. (2004). Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica*, *72*, 1295–1319.

Hanson, B. A. (2000). *IRT parameter estimation using the EM algorithm* (Tech. Rep.). http://www.b-a-h.com/.

Harding, M. (2012). *Estimating the number of factors in large dimensional factor models* (Tech. Rep.).

Hartig, J. (2008). Psychometric models for the assessment of competencies. In J. Hartig, E. Klieme & D. Leutner (Hrsg.), *Assessment of competencies in educational contexts* (S. 69–90). Göttingen: Hogrefe & Huber.

Hartig, J. & Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychological Test and Assessment Modeling*, *54*, 418–431.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*, 139–164.

Hayduk, L. A. (1987). *Structural equation modeling with LISREL: Essentials and advances.* JHU Press.

Hecht, M., Weirich, S., Siegle, T. & Frey, A. (2015). Modeling booklet effects for nonequivalent group designs in large-scale assessment. *Educational and Psychological Measurement*, *75*, 568–584.

Hedeker, D., Mermelstein, R. J. & Demirtas, H. (2008). An application of a mixed-effects location scale model for analysis of ecological momentary assessment (EMA) data. *Biometrics*, *64* (2), 627–634.

Heinen, T. (1993). *Discrete latent variable models.* Tilburg: Tilburg University Press.

Heyer, D. & Niederee, R. (1989). Elements of a model-theoretic framework for probabilistic measurement. In E. E. Roskam (Hrsg.), *Mathematical psychology in progress* (S. 99–112). Berlin: Springer.

Heyer, D. & Niederee, R. (1992). Generalizing the concept of binary choice systems induced by rankings: One way of probabilizing deterministic measurement structures. *Mathematical Social Sciences*, *23*, 31–44.

Hill, C. J., Bloom, H. S., Black, A. R. & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*, 172–177.

Ho, A. D. (2009). A nonparametric framework for comparing trends and gaps across tests. *Journal of Educational and Behavioral Statistics*, *34*, 201–228.

Hoeting, J. A., Madigan, D., Raftery, A. E. & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*, 382–417.

Hohensinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L. & Frebort, M. (2008). Examining item-position effects in large-scale assessments using the linear logistic test model. *Psychology Science Quarterly*, *50*, 391–402.

Hoijtink, H. & Molenaar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using the gibbs sampler and posterior predictive checks. *Psychometrika*, *62*(2), 171–189.

Hoijtink, H. & Vollema, M. (2003). Contemporary extensions of the Rasch model. *Quality & Quantity*, *37*, 263–276.

Holland, P. W. (1990a). On the sampling theory foundations of item response theory models. *Psychometrika*, *55*, 577–601.

Holland, P. W. (1990b). The Dutch identity: A new tool for the study of item response models. *Psychometrika*, *55*, 5–18.

Holland, P. W. & Wainer, H. (Hrsg.). (1993). *Differential item functioning: Theory and practice.* Hillsdale, NJ: Erlbaum.

Holman, R. & Glas, C. A. W. (2005). Modelling nonignorable missing data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, *58*, 1–17.

Hong, H., Wang, C., Lim, Y. S. & Douglas, J. (2015). Efficient models for cognitive diagnosis with continuous and mixed-type latent variables. *Applied Psychological Measurement*, *39*(1), 31–43.

Hosenfeld, I. (2008). Monitoring and assurance of school quality:Principles of assessment and internet-based feedback of test results. In J. Hartig, E. Klieme & D. Leutner (Hrsg.), *Assessment of competencies in educational contexts* (S. 337–356). Göttingen: Hogrefe & Huber.

Hoskens, M. & De Boeck, P. (1995). Componential IRT models for polytomous items. *Journal of Educational Measurement*, *32*, 364–384.

Hoskens, M. & De Boeck, P. (1997). A parametric model for local dependence among test items. *Psychological Methods*, *2*, 261–277.

Howell, R. D. (2014). What is the latent variable in causal indicator models? *Measurement: Interdisciplinary Research & Perspectives*, *12*(4), 141–145.

Hulin, C., Cudeck, R., Netemeyer, R., Dillon, W. R., McDonald, R. P. & Bearden, W. (2001). Measurement. *Journal of Consumer Psychology*, *10*, 55-69.

Hunter, J. E. (1968). Probabilistic foundations for coefficients of generalizability. *Psychometrika*, *33*, 1–18.

Husek, T. R. & Sirotnik, K. (1967). *Item sampling in educational research* (CSEIP Occasional Report No. 2). Los Angeles: University of California.

Hutchison, D. (2008). On the conceptualisation of measurement error. *Oxford Review of Education*, *34*, 443–460.

Ip, E. H. (2000). Adjusting for information inflation due to local dependency in moderately large item clusters. *Psychometrika*, *65*, 73–91.

Ip, E. H. (2010). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *British Journal of Mathematical and Statistical Psychology*, *63*, 395–416.

Ip, E. H. & Chen, S.-H. (2012). Projective item response model for test-independent measurement. *Applied Psychological Measurement*, *36*, 581–601.

Ip, E. H., Molenberghs, G., Chen, S.-H., Goegebeur, Y. & De Boeck, P. (2013). Functionally unidimensional item response models for multivariate binary data. *Multivariate Behavioral Research*, *48*, 534–562.

Irtel, H. (1995a). An extension of the concept of specific objectivity. *Psychometrika*, *60*, 115–118.

Irtel, H. (1995b). *Entscheidungs- und testtheoretische Grundlagen der Psychologischen Diagnostik.* Mannheim: Universität Mannheim.

Irtel, H. & Schmalhofer, F. (1982). Psychodiagnostik auf Ordinalskalenniveau: Messtheoretische Grundlagen, Modelltest und Parameterschätzung. *Archiv für Psychologie*, *134*, 197–218.

Jak, S., Oort, F. J. & Dolan, C. V. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling*, *20*(2), 265–282.

Jak, S., Oort, F. J. & Dolan, C. V. (2014). Measurement bias in multilevel data. *Structural Equation Modeling*, *21*(1), 31–39.

Jang, E. E. & Roussos, L. (2007). An investigation into the dimensionality of TOEFL using conditional covariance-based nonparametric approach. *Journal of Educational Measurement*, *44*, 1–21.

Jarvis, C., MacKenzie, S. & Podsakoff, P. A. (2003). Critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research*, *30*, 199–218.

Jiao, H., Kamata, A., Wang, S. & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement*, *49*(1), 82–100.

Jiao, H. & Zhang, Y. (2015). Polytomous multilevel testlet models for testlet-based assessments with complex sampling designs. *British Journal of Mathematical and Statistical Psychology*, *68*(1), 65–83.

Joe, H. (2015). *Dependence modeling with copulas.* Boca Raton: Chapman & Hall.

Johnson, M. S., Sinharay, S. & Bradlow, E. T. (2007). Hierarchical item response models. In C. R. Rao & S. Sinharay (Hrsg.), *Handbook of Statistics, Volume 26: Psychometrics* (S. 587–606). North Holland: Elsevier.

Johnson, V. E. & Albert, J. H. (1999). *Ordinal data modeling.* New York: Springer.

Junker, B. W. (1991). Essential independence and likelihood-based ability estimation for polytomous items. *Psychometrika*, *56*, 255–278.

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics*, *21*, 1359–1378.

Junker, B. W. & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258–272.

Kaiser, H. F. & Caffrey, J. (1965). Alpha factor analysis. *Psychometrika*, *30*, 1–14.

Kamata, A. & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*, *15*, 136–153.

Kamata, A. & Cheong, Y. F. (2007). Multilevel Rasch models. In M. von Davier & C. Carstensen (Hrsg.), *Multivariate and mixture distribution Rasch models* (S. 271–322). New York: Springer.

Kane, M. (1982). A sampling model for validity. *Applied Psychological Measurement*, *6*, 125–160.

Kane, M. (2002). Inferences about variance components and reliability-generalizability coefficients in the absence of random sampling. *Journal of Educational Measurement*, *39*, 165–181.

Kane, M. (2006). Validation. In R. L. Brennan (Hrsg.), *Educational measurement* (S. 17–64). Westport, CT: American Council on Education.

Kane, M. (2011). The errors of our ways. *Journal of Educational Measurement*, *48*, 12–30.

Kane, M. T., Gillmore, G. M. & Crooks, T. J. (1976). Student evaluations of teaching: The generalizability of class means. *Journal of Educational Measurement*, *13*(3), 171–183.

Kaplan, D. (2014). *Bayesian statistics for the social sciences.* New York: Guilfort Press.

Kass, R. E. (2011). Statistical inference: The big picture (with discussion). *Statistical Science*, *26*, 1–20.

Kennedy, M. C. & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(3), 425–464.

Khanna, R., Zhang, L., Agarwal, D. & Chen, B.-C. (2013). Parallel matrix factorization for binary response. In *2013 IEEE International Conference on Big Data* (S. 430-438).

Kiefer, T., Robitzsch, A. & Wu, M. (2015). *TAM: Test analysis modules.* (R package version 1.6-0)

Kim, J.-S. & Bolt, D. M. (2007). Estimating item response models using markov chain Monte Carlo methods. *Educational Measurement*, *26*, 38–51.

Klicpera, C. & Gasteiger-Klicpera, B. (1993). *Lesen und Schreiben – Entwicklung und Schwierigkeiten: Die Wiener Längsschnittuntersuchungen über die Entwicklung, den Verlauf und die Ursachen von Lese- und Schreibschwierigkeiten in der Pflichtschulzeit.* Bern: Huber.

Koenker, R. & Yoon, J. (2009). Parametric links for binary choice models: A Fisherian-Bayesian colloquy. *Journal of Econometrics*, *152*, 120–130.

Köhler, C., Pohl, S. & Carstensen, C. H. (2015). Taking the missing propensity into account when estimating competence scores evaluation of item response theory models for nonignorable omissions. *Educational and Psychological Measurement*, *75*, 850–874.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Hrsg.), *Educational measurement* (S. 155–186). Westport: Praeger Publisher.

Kolen, M. J. & Brennan, R. L. (2004). *Test equating, scaling, and linking.* New York: Springer.

Kolenikov, S. (2010). Resampling variance estimation for complex survey data. *The Stata Journal*, *10*, 165–199.

Krupskii, P. & Joe, H. (2013). Factor copula models for multivariate data. *Journal of Multivariate Analysis*, *120*, 85–101.

Kubinger, K. D. (2008). On the revival of the Rasch model-based LLTM: From constructing tests using item generating rules to measuring item administration effects. *Psychology Science Quarterly*, *50*, 311–327.

Kubinger, K. D. (2009). Applications of the linear logistic test model in psychometric research. *Educational Psychological Measurement*, *69*, 232–244.

Kubinger, K. D. & Draxler, C. (2007). Probleme bei der Testkonstruktion nach dem Rasch-Modell. *Diagnostica*, *53*, 131–143.

Kubinger, K. D., Rasch, D. & Yanagida, T. (2011). *Statistik in der Psychologie.* Göttingen: Hogrefe.

Kurz, K., Kratzmann, J. & von Maurice, J. (2007). *Die BiKS-Studie. Methodenbericht zur Stichprobenziehung.* http://psydok.sulb.uni-saarland.de/volltexte/2007/990/index.html.

Kyngdon, A. (2008). The Rasch model from the perspective of the representational theory of measurement. *Theory & Psychology*, *18*(1), 89–109.

Kyngdon, A. (2011). Plausible measurement analogies to some psychometric models of test performance. *British Journal of Mathematical and Statistical Psychology*, *64*(3), 478–497.

Lando, T. & Bertoli-Barsotti, L. (2014). A modified minimum divergence estimator: some preliminary results for the Rasch model. *Electronic Journal of Applied Statistical Analysis*, *7*(1), 37–57.

Le, L. T. (2009). Effects of item positions on their difficulty and discrimination – A study in PISA science data across test language and countries. In K. Shigemasu, A. Okada, T. Imaizumi & T. Hoshino (Hrsg.), *New trends in psychometrics* (S. 207–214). Tokyo: Universal.

Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach.* Hoboken: Wiley.

Lehmann, R., Peek, R. & Poerschke, J. (2006). *HAMLET 3-4. Hamburger Lesetest für 3. und 4. Klassen.* Göttingen: Hogrefe.

Lenhard, W. & Schneider, W. (2005). *ELFE 1-6: Ein Leseverständnistest für Erst- bis Sechstklässler.* Göttingen: Hogrefe.

Levin, A. T. & Williams, J. C. (2003). Robust monetary policy with competing reference models. *Journal of Monetary Economics*, *50*, 945–975.

Levine, R. L. & Hunter, J. E. (1971). Statistical and psychometric inference in principal components analysis. *Multivariate Behavioral Research*, *6*, 105–116.

Levy, R. (2011). Bayesian data-model fit assessment for structural equation modeling. *Structural Equation Modeling*, *18*(4), 663–685.

Lewis, C. (2007). Selected topics in classical test theory. In C. R. Rao & S. Sinharay (Hrsg.), *Handbook of Statistics, Volume 26: Psychometrics* (S. 29–43). North Holland: Elsevier.

Li, Y., Bolt, D. M. & Fu, J. (2006). A comparison of alternative testlet models. *Applied Psychological Measurement*, *30*, 3–21.

Linacre, J. M. (1989). *Many-faceted Rasch measurement.* Chicago: MESA Press.

Lindsay, B., Clogg, C. C. & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, *86*(413), 96–107.

Little, T. D. (2013). *Longitudinal structural equation modeling.* New York: Guilfort Press.

Loken, E. & Rullison, K. L. (2010). Estimation of a four-parameter item response model. *British Journal of Mathematical and Statistical Psychology*, *63*, 509–525.

Longford, N. T. (2003). An alternative to model selection in ordinary regression. *Statistics and Computing*, *13*, 67–80.

Longford, N. T. (2008). *Studying human populations.* New York: Springer.

Longford, N. T., Holland, P. W. & Thayer, D. T. (1993). Stability of the MH-DIF statistics across populations. In P. W. Holland & H. Wainer (Hrsg.), *Differential item functioning: Theory and practice* (S. 171–196). Hillsdale, NJ: Erlbaum.

Lord, F. M. (1955). Sampling fluctuations resulting from the sampling of test items. *Psychometrika*, *20*, 1–22.

Lord, F. M. (1965). A strong true-score theory, with applications. *Psychometrika*, *30*(3), 239–270.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, *39*, 247–264.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading: Addison-Wesley.

Lucke, J. F. (2005). "Rassling the hog": The influence of correlated item error on internal consistency, classical reliability, and congeneric reliability. *Applied Psychological Measurement*, *29*, 106–125.

Lüdtke, O. & Robitzsch, A. (2010). Umgang mit fehlenden Daten in der empirischen Bildungsforschung. In S. Maschke & L. Stecher (Hrsg.), *Enzyklopädie Erziehungswissenschaft Online. Fachgebiet Methoden der empirischen erziehungswissenschaftlichen Forschung, Quantitative Forschungsmethoden.* Weinheim: Juventa.

Lüdtke, O., Robitzsch, A., Kenny, D. A. & Trautwein, U. (2013). A general and flexible approach to estimating the social relations model using bayesian methods. *Psychological Methods*, *18*(1), 101–119.

Lüdtke, O., Robitzsch, A., Trautwein, U. & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings in multilevel modelling. *Contemporary Educational Psychology*, *34*, 120–131.

Lunn, D., Spiegelhalter, D., Thomas, A. & Best, N. (2009). The BUGS project: evolution, critique and future directions. *Statistics in Medicine*, *28*(25), 3049–3067.

Lunn, D. J., Thomas, A., Best, N. & Spiegelhalter, D. (2000). WinBUGS – A Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, *10*(4), 325–337.

MacCallum, R. C. (2003). Working with imperfect models. *Multivariate Behavioral Research*, *38*, 113–139.

MacCallum, R. C., Browne, M. W. & Cai, L. (2007). Factor analysis models as approximations. In R. Cudeck & R. C. MacCallum (Hrsg.), *Factor analysis at 100* (S. 153–175). Mahwah, NJ: Lawrence Erlbaum.

MacCallum, R. C. & O'Hagan, A. (2015). Advances in modeling model discrepancy: Comment on Wu and Browne (2015). *Psychometrika*, *80*, 601–607.

MacCallum, R. C. & Tucker, L. R. (1991). Representing sources of error in the common-factor model: Implications for theory and practice. *Psychological Bulletin*, *109*(3), 502–511.

Magis, D., Tuerlinckx, F. & De Boeck, P. (2015). Detection of differential item functioning using the lasso approach. *Journal of Educational and Behavioral Statistics*, *40*(2), 111–135.

Marais, I. (2009). Response dependence and the measurement of change. *Journal of Applied Measurement*, *10*, 17–29.

Maris, G. & Bechger, T. (2009). On interpreting the model parameters for the three parameter logistic model. *Measurement: Interdisciplinary Research & Perspective*, *7*, 75–88.

Markus, K. A. & Borsboom, D. (2013a). *Frontiers of test validity theory: Measurement, causation, and meaning.* New York: Routledge.

Markus, K. A. & Borsboom, D. (2013b). Reflective measurement models, behavior domains, and common causes. *New Ideas in Psychology*, *31*, 54–64.

Maronna, R. A., Martin, R. D. & Yohai, V. J. (2006). *Robust statistics.* New York: Wiley.

Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics*, *31*, 35–62.

Maxwell, S. E. & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective.* New York: Psychology Press.

Maydeu-Olivares, A. (2005). Linear item response theory, nonlinear item response theory and factor analysis: A unified framework. In A. Maydeu-Olivares & J. J. McArdle (Hrsg.), *Contemporary psychometrics. A festschrift for Roderick P. McDonald.* (S. 73–102). New Jersey: Lawrence Erlbaum.

Maydeu-Olivares, A. & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, *49*(4), 305–328.

Mazzeo, J. & von Davier, M. (2008). *Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results.* EDU/PISA/GB(2008)28.

McCrae, R. R. (2015). A more nuanced view of reliability specificity in the trait hierarchy. *Personality and Social Psychology Review*, *19*(2), 97–112.

McDonald, R. P. (1978). Generalizability in factorable domains: „Domain validity and generalizability". *Educational and Psychological Measurement*, *38*(1), 75–79.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, *34*, 100–117.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. J. van der Linden & R. K. Hambleton (Hrsg.), *Handbook of modern item response theory* (S. 257–269). New York: Springer.

McDonald, R. P. (1999). *Test theory.* Mahwah NJ: Erlbaum.

McDonald, R. P. (2003). Behavior domains in theory and in practice. *Alberta Journal of Educational Research*, *49*, 212–230.

McDonald, R. P. (2010). Structural models and the art of approximation. *Perspectives on Psychological Science*, *5*(6), 675–686.

McDonald, R. P. (2011). Measuring latent quantities. *Psychometrika*, *76*, 511–536.

McDonald, R. P. & Mulaik, S. A. (1979). Determinacy of common factors: A nontechnical review. *Psychological Bulletin*, *86*, 297–306.

Meijer, R. R. & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107–135.

Meister, A. (2006). Density estimation with normal measurement error with unknown variance. *Statistica Sinica*, *16*(1), 195–211.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*(2), 127–143.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543.

Mevik, B.-H. & Wehrens, R. (2007). The pls package: Principal component and partial least squares regression in R. *Journal of Statistical Software*, *18*(2), 1–24.

Michaelides, M. P. (2010). A review of the effects on IRT item parameter estimates with a focus on misbehaving common items in test equating. *Frontiers in Psychology / Quantitative Psychology and Measurement*, *1*, 167.

Michaelides, M. P. & Haertel, E. H. (2004). *Sampling of common items: an unrecognized source of error in test equating* (Tech. Rep.). New York: Center for the Study of Evaluation and National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Michailidis, G. & de Leeuw, J. (1998). The Gifi system of descriptive multivariate analysis. *Statistical Science*, *13*, 307–336.

Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data: An introduction. *Child Development Perspectives*, *4*, 5–9.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance.* New York: Routledge.

Min, S. & He, L. (2014). Applying unidimensional and multidimensional item response theory models in testlet-based reading assessment. *Language Testing*, *31*, 453-477.

Minnamaier, G. (2002). Wie verläuft die Kompetenzentwicklung – kontinuierlich oder diskontinuierlich? In *4. BIBB-Fachkongress 2002.*

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*, 177–196.

Mislevy, R. J. & Wu, P.-K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (ETS Research Report RR96-30). Princeton: ETS.

Molenaar, D. (2015). Heteroscedastic latent trait models for dichotomous data. *Psychometrika*, *80*, 625–644.

Molenaar, D. & Borsboom, D. (2013). The formalization of fairness: issues in testing for measurement invariance using subtest scores. *Educational Research and Evaluation*, *19*(2-3), 223–244.

Molenaar, I. W. (1995). Some background for item response theory and the Rasch model. In G. H. Fischer & I. W. Molenaar (Hrsg.), *Rasch models. foundations, recent developments, and applications* (S. 3–14). New York: Springer.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research & Perspective*, *2*, 201–218.

Molenberghs, G. & Verbeke, G. (2005). *Models for discrete longitudinal data.* New York: Springer.

Monseur, C., Baye, A., Lafontaine, D. & Quittre, V. (2011). PISA test format assessment and the local independence assumption. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessment*, *4*, 131–155.

Monseur, C. & Berezner, A. (2007). The computation of equating errors in international surveys in education. *Journal of Applied Measurement*, *8*, 323–335.

Monseur, C., Sibberns, H. & Hastedt, D. (2008). Linking errors in trend estmation for international surveys in education. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *1*, 113–122.

Montgomery, J. M. & Nyhan, B. (2010). Bayesian model averaging: Theoretical developments and practical applications. *Political Analysis*, *18*, 245–270.

Morgan, S. L. & Winship, C. (2007). *Counterfactuals and causal inference.* Cambridge: Cambridge University Press.

Mulaik, S. A. (2009a). *Foundations of factor analysis.* Chapman & Hall.

Mulaik, S. A. (2009b). *Linear causal modeling with structural equations.* CRC Press.

Münnich, R. S., R. (2005). PRIMA: A new multiple imputation procedure for binary variables. *Journal of Official Statistics*, *21*, 325–341.

Muthén, B. & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological methods*, *17*(3), 313–335.

Muthén, B. & Asparouhov, T. (2014). IRT studies of many groups: the alignment method. *Frontiers in Psychology / Quantitative Psychology and Measurement*, *5*.

Muthén, B. O., Kao, C.-F. & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement*, *28*, 1–22.

Muthén, L. K. & Muthén, B. O. (1998-2010). *Mplus user's guide. Sixth edition.* Los Angeles.

Muthén, L. K. & Muthén, B. O. (1998-2013). *Mplus user's guide. Seventh edition.* Los Angeles.

Naumann, A., Hochweber, J. & Hartig, J. (2014). Modeling instructional sensitivity using a longitudinal multilevel differential item functioning approach. *Journal of Educational Measurement*, *51*(4), 381–399.

Niederée, R. & Mausfeld, R. (1996a). Das Bedeutsamkeitsproblem in der Statistik. In E. Erdfelder, R. Mausfeld, T. Meiser & R. Rudinger (Hrsg.), *Handbuch quantitative methoden* (S. 399–410). Weinheim: Psychologie Verlags Union.

Niederée, R. & Mausfeld, R. (1996b). Skalenniveau, Invarianz und „Bedeutsamkeit". In E. Erdfelder, R. Mausfeld, T. Meiser & R. Rudinger (Hrsg.), *Handbuch quantitative methoden* (S. 385–398). Weinheim: Psychologie Verlags Union.

Nikoloulopoulos, A. K. & Joe, H. (2015). Factor copula models for item response data. *Psychometrika*, *80*(1), 126–150.

Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric theory.* New York: McGraw-Hill.

Oberski, D. L. (2014). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis*, *22*(1), 45–60.

Oberski, D. L. & Satorra, A. (2013). Measurement error models with uncertainty about the error variance. *Structural Equation Modeling*, *20*(3), 409–428.

OECD. (2005). *PISA 2003: Technical report.* Paris.

OECD. (2014). *PISA 2012: Technical report.* Paris.

Ogasawara, H. (2002). Stable response functions with unstable item parameter estimates. *Applied Psychological Measurement*, *26*, 239–254.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics*, *168*(2), 244–258.

Park, T. & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686.

Patz, R. J. & Junker, B. W. (1999a). A straighforward approach to Markov chain Monte Carlo methods. *Journal of Educational and Behavioral Statistics*, *24*, 146–178.

Patz, R. J. & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*, 342–366.

Paul, D. & Aue, A. (2014). Random matrix theory in statistics: a review. *Journal of Statistical Planning and Inference*, *150*, 1–29.

Peress, M. (2012). Identification of a semiparametric item response model. *Psychometrika*, *77*, 223–243.

Perline, R., Wright, B. D. & Wainer, H. (1979). The rasch model as additive conjoint measurement. *Applied Psychological Measurement*, *3*(2), 237–255.

Peters, G.-J. Y. (2014). The alpha and the omega of scale reliability and validity: Why and how to abandon Cronbach's alpha and the route towards more comprehensive assessment of scale quality. *European Health Psychologist*, *16*(2), 56–69.

Pfanzagl, J. (1968). *Theory of measurement.* Würzburg: Physica-Verlag.

Pfost, M., Karing, C., Lorenz, C. & Artelt, C. (2010). Schereneffekte im ein- und mehrgliedrigen Schulsystem. Differenzielle Entwicklung sprachlicher Kompetenzen am Übergang von der Grund- in die weiterführende Schule? *Zeitschrift für Pädagogische Psychologie*, *24*, 259–273.

Pietsch, M. (2011). Fehlende Daten bei Unterrichtsbeobachtungen: Eine Sensitivitätsanalyse anhand von Daten der Schulinspektion Hamburg. *Empirische Pädagogik*, *25*, 47–87.

Pohl, S. & Carstensen, C. (2012). *NEPS technical report - Scaling the data of the competence tests* (NEPS Working Paper No. 14). Bamberg: NEPS.

Pohl, S. & Carstensen, C. H. (2013). Scaling of competence tests in the National Educational Panel Study: Many questions, some answers, and further challenges. *Journal for Educational Research Online*, *5*, 189–216.

Pohl, S., Gräfe, S. & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement*, *74*, 423–452.

Polikoff, M. S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice*, *29*(4), 3–14.

Pornprasertmanit, S., Lee, J. & Preacher, K. J. (2014). Ignoring clustering in confirmatory factor analysis: Some consequences for model fit and standardized parameter estimates. *Multivariate Behavioral Research*, *49*(6), 518–543.

Pourahmadi, M. (2013). *High-dimensional covariance estimation.* Hoboken: Wiley.

Press, S. J. (2003). *Subjective and objective Bayesian statistics.* Hoboken: Wiley.

Press, S. J. & Shigemasu, K. (1997). *Bayesian inference in factor analysis (revised)* (Tech. Rep.).

Proctor, C. H. (1970). A probabilistic formulation and statistical analysis of Guttman scaling. *Psychometrika*, *35*(1), 73–78.

R Core Team. (2014). *R: A language and environment for statistical computing.* Vienna, Austria.

Raiche, G., Magis, D., Blais, J.-G. & Brochu, P. (2013). Taking atypical response patterns into account. In M. Simou, K. Ercikan & M. Rousseau (Hrsg.), *Improving large scale education assessment: Theory, issues, and practice* (S. 238–259). New York: Routledge.

Rajaratnam, N., Cronbach, L. J. & Gleser, G. C. (1965). Generalizability of stratified-parallel tests. *Psychometrika*, *30*, 39–56.

Ramsay, J. O. (1989). A comparison of three simple test theory models. *Psychometrika*, *54*, 487–499.

Ramsay, J. O. (1996). A geometric approach to item response theory. *Behaviormetrika*, *23*, 3–16.

Ramsay, J. O. & Silverman, B. W. (2005). *Functional data analysis.* New York: Springer.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. In *Danish yearbook of philosophy* (S. 58–93). Copenhagen: Munksgaard. (URL http://www.rasch.org/memo18.htm)

Rauch, W. A. & Moosbrugger, H. (2011). Klassische Testtheorie. Grundlagen und Erweiterungen für heterogene Tests und Mehrfacettenmodelle. In L. F. Hornke, M. Amelang & M. Kersting (Hrsg.), *Enzyklopädie der Psychologie: Themenbereich B Methodologie und Methoden, Serie II Psychologische Diagnostik, Band 2, Methoden der psychologischen Diagnostik* (S. 1–87). Göttingen: Hogrefe.

Raykov, T. & Marcoulides, G. A. (2010). *Introduction to psychometric theory.* Taylor & Francis.

Reckase, M. D. (2009). *Multidimensional item response theory.* New York: Springer.

Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research*, *35*(4), 543–568.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*, 667–696.

Reise, S. P., Moore, T. M. & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, *92*, 544–559.

Renard, D., Molenberghs, G. & Geys, H. (2004). A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics & Data Analysis*, *44*, 649–667.

Resseguier, N., Roch, G. & Paoletti, X. (2011). Sensitivity analysis: When data are missing not-at-random. *Epidemiology*, *22*, 282–283.

Retelsdorf, J. & Möller, J. (2008). Entwicklungen von Lesekompetenz und Lesemotivation. Schereneffekte in der Sekundarstufe? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *40*, 179–188.

Revelle, W. (2015). *An introduction to psychometric theory with applications in R.* http://personality-project.org/r/book/.

Revelle, W. & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*, 145–154.

Rhemtulla, M., Bork, R. van & Borsboom, D. (2015). Calling models with causal indicators „measurement models“ implies more than they can deliver. *Measurement: Interdisciplinary Research & Perspectives*, *13*(1), 59–62.

Richter, T. & Christmann, U. (2002). Lesekompetenz: Prozessebenen und interindividuelle Unterschiede. In N. Groeben & B. Hurrelmann (Hrsg.), *Lesekompetenz: Bedingungen, Dimensionen, Funktionen* (S. 25–58). Weinheim: Juventa.

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, *47*(3), 361–372.

Robitzsch, A. (2009). Methodische Herausforderungen bei der Kalibrierung von Leistungstests. In A. Bremerich-Vos, D. Granzer & O. Köller (Hrsg.), *Bildungsstandards Deutsch und Mathematik* (S. 42–106). Weinheim: Beltz.

Robitzsch, A. (2010). TIMSS 1995 und 2007: Trend der mathematischen Kompetenzen in Österreich. In B. Suchán, C. Wallner-Paschon & C. Schreiner (Hrsg.), *TIMSS 2007. Österreichischer Expertenbericht* (S. 56–63). Graz: Leykam.

Robitzsch, A. (2011a). *Berechnung von Standardfehlern in Large-Scale Assessments auf Grund von Item Sampling für Ländermittelwerte sowie originale und marginale Trendschätzungen.* 10. Tagung der Fachgruppe Methoden und Evaluation, Bamberg.

Robitzsch, A. (2011b). *Multilevel-DIF-Modelle: Bedeutung differenziellen Itemfunktionierens zwischen Schulklassen.* 75. AEPF-Tagung, Bamberg.

Robitzsch, A. (2012). *Wie signifikant ist der österreichische „Absturz“ im Lesen in der PISA-Studie 2009? Eine Erklärung mittels alternativer Berechnungen von Standardfehlern für Ländermittelwerte sowie für originale und marginale Trendschätzungen.* 77. AEPF-Tagung, Bielefeld.

Robitzsch, A. (2015). *sirt: Supplementary item response theory models.* (R package version 1.5-0)

Robitzsch, A., Dörfler, T., Pfost, M. & Artelt, C. (2011). Die Bedeutung der Itemauswahl und der Modellwahl für die längsschnittliche Erfassung von Kompetenzen: Lesekompetenzentwicklung in der Primarstufe. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *43*, 213–227.

Robitzsch, A., Freunberger, R., Itzlinger-Bruneforth, U., Breit, S. & Schreiner, C. (2015). *Ein Kommentar zu Vohns: „Bildungsstandards M8 - Wie kommen die offiziellen Zahlen zustande und was sagen sie (nicht) aus?“* (Tech. Rep.). Salzburg: BIFIE. URL: https://www.bifie.at/node/2842.

Robitzsch, A., Kiefer, T., George, A. C. & Ünlü, A. (2015). *CDM: Cognitive diagnosis modeling.* (R package version 4.2-12)

Robitzsch, A. & Lüdtke, O. (2015). Kommentar zum Beitrag „Lokale Abhängigkeit von Items im TestDaF-Leseverstehen“ von Thomas Eckes. *Diagnostica*, *61*, 107–109.

Rohwer, G. (2013). *Making sense of missing answers in competence tests* (NEPS Working Paper No. 30). Bamberg: NEPS.

Rohwer, G. & BloSSfeld, H.-P. (2012). *Contextual and random coefficient multilevel models. A comparison* (NEPS Working Paper No. 6). Bamberg: NEPS.

Rose, N. (2013). *Item nonresponses in educational and psychological assessment.* Friedrich-Schiller-Universität Jena, Dissertationsschrift.

Rose, N., von Davier, M. & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT)* (ETS Research Report RR10-11). Princeton: ETS.

Rossi, G. B. (2006). A probabilistic theory of measurement. *Measurement*, *39*, 34–50.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion.* Bern: Huber.

Roussos, L. & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, *20*, 355–371.

Rowe, D. B. (2002). *Multivariate Bayesian statistics: Models for source separation and signal unmixing.* Boca Raton: CRC Press.

Rupp, A. A., Dey, D. K. & Zumbo, B. D. (2004). To Bayes or not to Bayes, from whether to when: Applications of Bayesian methodology to modeling. *Structural Equation Modeling*, *11*, 424–451.

Rupp, A. A. & Templin, J. (2008). The effects of Q-Matrix misspecifications on parameter estimates and classification accuracy in the DINA model. *Educational Psychological Measurement*, *68*, 78–96.

Rupp, A. A. & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research & Perspective*, *6*(4), 219–262.

Rupp, A. A. & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement*, *66*, 63–84.

Rutkowski, L., Gonzales, E., von Davier, M. & Zhou, Y. (2014). Assessment design for international large-scale assessments. In L. Rutkowski, M. von Davier & D. Rutkowski (Hrsg.), *Handbook of international large-scale assessment* (S. 75-95). Boca Raton: CRC Press.

Samejima, F. (1997). Graded response model. In G. H. Fischer & I. W. Molenaar (Hrsg.), *Handbook of modern item response theory* (S. 85–100). New York: Springer.

Samejima, F. (2010). The general graded response model. In M. L. Nering & R. Ostini (Hrsg.), *Handbook of polytomous item response models* (S. 43–76). New York: Routledge.

San Martin, E. & De Boeck, P. (2015). What do you mean by a difficult item? On the interpretation of the difficulty parameter in a Rasch model. In R. E. Millsap, D. M. Bolt, L. van der Ark & W.-C. Wang (Hrsg.), *Quantitative psychology research. springer proceedings in mathematics & statistics 89* (S. 1–14). New York: Springer.

San Martin, E., Del Pino, G. & De Boeck, P. (2006). IRT models for ability-based guessing. *Applied Psychological Measurement*, *30*(3), 183–203.

San Martin, E., González, J. & Tuerlinckx, F. (2009). Identified parameters, parameters of interest and their relationships. *Measurement: Interdisciplinary Research & Perspective*, *7*, 97–105.

San Martin, E. & Rolin, J.-M. (2013). Identification of parametric Rasch-type models. *Journal of Statistical Planning and Inference*, *143*, 116–130.

San Martin, E., Rolin, J.-M. & Castro, L. M. (2013). Identification of the 1PL model with guessing parameter: parametric and semi-parametric results. *Psychometrika*, 341–379.

San Martin, E., González, J. & Tuerlinckx, F. (2015). On the unidentifiability of the fixed-effects 3PL model. *Psychometrika*, *80*, 450–467.

Saris, W. E., Satorra, A. & Van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling*, *16*(4), 561–582.

Särndal, C.-E., Swensson, B. & Wretman, J. (1992). *Model assisted survey sampling.* New York: Springer.

Satorra, A. (2015). A comment on a paper by H. Wu and M. W. Browne. *Psychometrika*, *80*, 613–618.

Savalei, V. (2014). Understanding robust corrections in structural equation modeling. *Structural Equation Modeling*, *21*, 149–160.

Scheerer-Neumann, G. (1997). Lesen und Leseschwierigkeiten. In F. E. Weinert (Hrsg.), *Psychologie des Unterrichts und der Schule* (S. 279–325). Göttingen: Hogrefe.

Scheiblechner, H. (1995). Isotonic ordinal probabilistic models (ISOP). *Psychometrika*, *60*(2), 281–304.

Scheiblechner, H. (1999). Additive conjoint isotonic probabilistic models (ADISOP). *Psychometrika*, *64*(3), 295–316.

Scheiblechner, H. (2007). A unified nonparametric IRT model for d-dimensional psychological test data (d-ISOP). *Psychometrika*, *72*, 43–67.

Schmittmann, V. D., Cramer, A. O. J., Waldorp, L. J., Epskamp, S., Kievit, R. A. & Borsboom, D. (2013). Deconstructing the construct: A network perspective on psychological phenomena. *New Ideas in Psychology*, *31*, 43–53.

Schneider, W. & Stefanek, J. (2004). Entwicklungsveränderungen allgemeiner kognitiver Fähigkeiten und schulbezogener Fertigkeiten im Kindes- und Jugendalter. Evidenz für einen Schereneffekt? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *36*, 147–159.

Schroeders, U., Robitzsch, A. & Schipolowski, S. (2014). A comparison of different psychometric approaches to modeling testlet structures: An example with c-tests. *Journal of Educational Measurement*, *51*(4), 400–418.

Schulz, E. M. & Nicewander, W. A. (1997). Grade equivalent and IRT representation of growth. *Journal of Educational Measurement*, *34*, 315–331.

Schwabe, F. & Gebauer, M. M. (2013). (Test-)Fairness – eine Herausforderung an standardisierte Leistungsdiagnostik. In N. McElvany, M. M. Gebauer, W. Bos & H. G. Holtappels (Hrsg.), *Sprachliche, kulturelle und soziale Heterogenität in der Schule als Herausforderung und Chance der Schulentwicklung. IFS-Jahrbuch der Schulentwicklung, Bd. 17* (S. 217–236). Weinheim: Juventa.

Schwantner, U. & Schreiner, C. (Hrsg.). (2010). *PISA 2009. Erste Ergebnisse.* Graz: Leykam.

Searle, S. R., Casella, G. & McCulloch, C. E. (1992). *Variance components.* New York: John Wiley & Sons.

Segawa, E., Emery, S. & Curry, S. J. (2008). Extended generalized linear latent and mixed model. *Journal of Educational and Behavioral Statistics*, *33*, 464–488.

Shojima, K. (2007). *Maximum likelihood estimation of latent rank under neural test model* (Tech. Rep.). DThe National Center for University Entrance Examinations.

Si, Y. & Reiter, J. P. (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, *38*(5), 499–521.

Siddique, J., Harel, O. & Crespi, C. M. (2012). Addressing missing data mechanism uncertainty using multiple-model multiple imputation: Application to a longitudinal clinical trial. *Annals of Applied Statistics*, *6*, 1814–1837.

Siemsen, E. & Bollen, K. A. (2007). Least absolute deviation estimation in structural equation modeling. *Sociological Methods & Research*, *36*, 227–265.

Sijtsma, K. (2006). Psychometrics in psychological research: Role model or partner in science? *Psychometrika*, *71*, 451–455.

Sijtsma, K. & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory.* Thousand Oaks: Sage.

Sijtsma, K. & van der Ark, L. A. (2015). Conceptions of reliability revisited and practical recommendations. *Nursing Research*, *64*(2), 128–136.

Sinharay, S. (2004). Experiences with Markov chain Monte Carlo convergence assessment in two psychometric examples. *Journal of Educational and Behavioral Statistics*, *29*, 461–488.

Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling.* Boca Raton: Chapman & Hall.

Snijders, T. A. B. & Bosker, R. J. (2012). *Multilevel analysis.* London: Sage.

Song, W., Yao, W. & Xing, Y. (2014). Robust mixture regression model fitting by Laplace distribution. *Computational Statistics & Data Analysis*, *71*, 128–137.

Song, X.-Y. & Lee, S.-Y. (2012). *Basic and advanced Bayesian structural equation modeling.* Hoboken: Wiley.

Spiegelhalter, D., Thomas, A., Best, N. & Lunn, D. (2003). *WinBUGS version 1.4 user manual.* MRC Biostatistics Unit.

Spiess, M. & Hamerle, A. (1996). On the properties of GEE estimators in the presence of invariant covariates. *Biometrical Journal*, *38*, 931–940.

Spiess, M., Nagl, W. & Hamerle, A. (1997). *Probit models: Regression parameter estimation using the ML principle despite misspecification of the correlation structure* (Research Report, Sonderforschungsbereich 386, Paper 67). LMU Munich.

Spoden, C., Fleischer, J. & Leutner, D. (2014). Niedrige Testmodellpassung als Resultat mangelnder Auswertungsobjektivität bei der Kodierung landesweiter Vergleichsarbeiten durch Lehrkräfte. *Journal für Mathematik-Didaktik*, *35*(1), 79–99.

Spray, J. A., Davey, T. C., Reckase, M. D., Ackerman, T. A. & Carlson, J. E. (1990). *Comparison of two logistic multidimensional item response theory models* (ACT Research Report No. ACT-RR-ONR-90-8). ACT.

Stenner, A. J., Burdick, D. S. & Stone, M. H. (2008). Formative and reflective models: Can a Rasch analysis tell the difference? *Rasch Measurement Transactions*, *22*, 1152–1153.

Steyer, R. (1989). Models of classical psychometric test theory as stochastic measurement models: Representation, uniqueness, meaningfulness, identifiability, and testability. *Methodika*, *3*, 25–60.

Steyer, R. & Eid, M. (2001). *Messen und Testen.* Berlin: Springer.

Steyer, R., Sengewald, E. & Hahn, S. (2015). Some comments on Wu and Browne. *Psychometrika*, *80*, 608–610.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589–617.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, *55*, 293–325.

Strandmark, N. L. & Linn, R. L. (1987). A generalized logistic item response model parameterizing test score inappropriateness. *Applied psychological measurement*, *11*(4), 355–370.

Strathmann, A. M. & Klauer, K. J. (2010). Lernverlaufsdiagnostik: Ein Ansatz zur längerfristigen Lernfortschrittsmessung. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *42*, 111–122.

Strobl, C. (2010). *Das Rasch-Modell.* München: Rainer Hampp Verlag.

Stukel, T. A. (1988). Generalized logistic models. *Journal of the American Statistical Association, 83*, 426–431.

Suh, Y. & Bolt, D. M. (2010). Nested logit models for multiple-choice item response data. *Psychometrika, 75*(3), 454–473.

Swygert, K. A., McLeod, L. D. & Thissen, D. (2001). Item response theory applied to combinations of multiple-choice and constructed response items - scale scores for patterns of sum scores. In D. Thissen & H. Wainer (Hrsg.), *Test scoring* (S. 217–250). Hillsdale, NJ: Erlbaum.

Takane, Y. & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*(3), 393–408.

te Marvelde, J. M., Glas, C. A., Van Landeghem, G. & Van Damme, J. (2006). Application of multidimensional item response theory models to longitudinal data. *Educational and Psychological Measurement, 66*(1), 5–34.

ten Berge, J. M. (1993). *Least squares optimization in multivariate analysis.* DSWO Press, Leiden University Leiden.

Thomas, D. R. & Cyr, A. (2002). Applying item response theory methods to complex survey data. In *Proceedings of the survey methods section* (S. 17–25). SSC Annual Meeting, May 2002 Proceedings of the Survey Methods Section.

Tryon, R. C. (1957). Reliability and behavior domain validity: Reformulation and historical critique. *Psychological Bulletin, 54*, 229–249.

Tuerlinckx, F. & De Boeck, P. (2004). Models for residual dependencies. In P. De Boeck & M. Wilson (Hrsg.), *Explanatory item response models: A generalized linear and nonlinear approach* (S. 289–316). New York: Springer.

Tuerlinckx, F. & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods, 6*, 181–195.

Tutz, G. (1989). *Latent Trait-Modell für ordinale Beobachtungen.* Berlin: Springer.

Tutz, G. (2012). *Regression for categorical data.* Cambridge: Cambridge University Press.

Tutz, G. & Schauberger, G. (2015). A penalty approach to differential item functioning in rasch models. *Psychometrika, 80*, 21–43.

van Borkulo, C. D., Borsboom, D., Epskamp, S., Blanken, T. F., Boschloo, L., Schoevers, R. A. & Waldorp, L. J. (2014). A new method for constructing networks from binary data. *Scientific Reports, 4*(5918).

van Buuren, S. (2012). *Flexible imputation of missing data.* Boca Raton: Chapman & Hall.

van Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45(3).*

Van De Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J. & Muthén, B. (2013). Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology | Quantitative Psychology and Measurement*, *4*.

van den Berg, S. M., Glas, C. A. W. & Boomsma, A. (2007). Variance decomposition using an IRT measurement model. *Behavior Genetics*, *37*, 604–616.

van den Heuvel-Panhuizen, M., Robitzsch, A., Treffers, A. & Köller, O. (2009). Large-scale assessment of change in student achievement: Dutch primary school students' results on written division in 1997 and 2004 as an example. *Psychometrika*, *74*, 351–365.

Van den Noortgate, W. & De Boeck, P. (2005). Assessing and explaining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics*, *30*(4), 443–464.

Van den Noortgate, W., De Boeck, P. & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics*, *28*, 369–386.

van der Leeden, R., Meijer, E. & Busing, F. M. T. A. (2007). Resampling multilevel models. In J. de Leeuw & E. Meijer (Hrsg.), *Handbook of multilevel analysis* (S. 401–434). New York: Springer.

van der Linden, W. J. (1979). Binomial test models and item difficulty. *Applied Psychological Measurement*, *3*, 401–411.

van der Linden, W. J. (1994). Fundamental measurement and the fundamentals of Rasch measurement. In M. Wilson (Hrsg.), *Objective Measurement: Theory into Practice Vol. 2* (S. 3–24). Norwod, NJ: Ablex Publishing Cooperation.

van der Linden, W. J. (2001). Book review: Applying the Rasch model by Bond & Fox. *International Journal of Testing*, *1*, 319–326.

van der Maas, H. J. L., Molenaar, D., Maris, G., Kievit, R. A. & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, *318*, 339–356.

VanderWeele, T. J. (2010). Direct and indirect effects for neighborhood-based clustered and longitudinal data. *Sociological Methods & Research*, *38*(4), 515–544.

Verhelst, N. D., Glas, C. A. W. & Verstralen, H. H. F. M. (1995). *One Parameter Logistic Model (OPLM)*. CITO, Arnhem.

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology*, *33*(1), 213–239.

Vermunt, J. K. (2008). Multilevel latent variable modeling: An application in educational testing. *Austrian Journal of Statistics*, *37*, 285–299.

Vermunt, J. K. & Magidson, J. (2005). Hierarchical mixture models for nested data structures. In C. Weihs & W. Gaul (Hrsg.), *Classification: The ubiquitous challenge* (S. 240–247). Heidelberg: Springer.

Viertl, R. (2006). Univariate statistical analysis with fuzzy data. *Computational Statistics & Data Analysis*, *51*(1), 133–147.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 287–307.

von Davier, M. (2009). Some notes on the reinvention of latent structure models as diagnostic classification models. *Measurement: Interdisciplinary Research & Perspective*, *7*(1), 67–74.

von Davier, M. (2010). Why sum scores may not tell us all about test takers. *Newborn and Infant Nursing Reviews*, *10*, 27–36.

von Davier, M. & Sinharay, S. (2014). Analytics in international large-scale assessments: Item response theory and population models. In L. Rutkowski, M. von Davier & D. Rutkowski (Hrsg.), *Handbook of international large-scale assessment* (S. 155–174). Boca Raton: CRC Press.

von Davier, M. & von Davier, A. A. (2007). A unified approach to IRT scale linking and scale transformations. *Methodology*, *3*, 115–124.

Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement*, *30*, 1–21.

Wainer, H. (2010a). 14 conversations about three things. *Journal of Educational and Behavioral Statistics*, *35*, 5–25.

Wainer, H. (2010b). Schrödinger's cat and the conception of probability in item response theory. *Chance*, *23*, 53–56.

Wainer, H., Bradlow, E. T. & Wang, X. (2007). *Testlet response theory and its applications.* Cambridge: Cambridge University Press.

Wainer, H. & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice*, *15*, 22–29.

Walther, G. & Granzer, D. (2009). Kompetenzmodell Mathematik. In A. Bremerich-Vos, D. Granzer & O. Köller (Hrsg.), *Bildungsstandards Deutsch und Mathematik* (S. 112–124). Weinheim: Beltz.

Wang, C. & Nydick, S. W. (2015). Comparing two algorithms for calibrating the restricted non-compensatory multidimensional IRT model. *Applied Psychological Measurement*, *39*(2), 119–134.

Wang, W.-C. & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*, 126–149.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450.

Warner, R. M., Kenny, D. A. & Stoto, M. (1979). A new round robin analysis of variance for social interaction data. *Journal of Personality and Social Psychology*, *37*(10), 1742–1757.

Wasserman, L. (2004). *All of statistics.* New York: Springer.

Weigel, A. P., Knutti, R., Liniger, M. A. & Appenzeller, C. (2010). Risks of model weighting in multimodel climate projections. *Journal of Climate*, *23*, 4175–4191.

Weirich, S., Hecht, M. & Böhme, K. (2014). Modeling item position effects using generalized linear mixed models. *Applied Psychological Measurement*, *38*(7), 535–548.

Weirich, S., Penk, C., Hecht, M., Roppelt, A. & Böhme, K. (submitted). Item position effects are moderated by changes in test-taking effort. *Journal of Educational Measurement*, *xx*, xxx–xxx.

Westfall, P. H., Henning, K. S. & Howell, R. D. (2012). The effect of error correlation on interfactor correlation in psychometric measurement. *Structural Equation Modeling*, *19*(1), 99–117.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1–25.

Williams, J. S. (1978). A definition for the common-factor analysis model and the elimination of problems of factor score indeterminacy. *Psychometrika*, *43*, 293–306.

Williams, J. S. (1979). A Biometrics invited paper. A synthetic basis for a comprehensive factor-analysis theory. *Biometrics*, 719–733.

Williams, L. J. & O'Boyle, E. (2011). The myth of global fit indices and alternatives for assessing latent variable relations. *Organizational Research Methods*, *14*, 350–369.

Willmott, A. S. & Fowles, D. F. (1974). *The objective interpretation of test performance* (Slough, NFER).

Wilson, M. (2012). Responding to a challenge that learning progressions pose to measurement practice. In A. C. Alonzo & A. W. Gotwals (Hrsg.), *Learning progressions in science* (S. 317–343). New York: Springer.

Wilson, M. (2013). Seeking a balance between the statistical and scientific elements in psychometrics. *Psychometrika*, *78*(2), 211–236.

Wilson, M. & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics*, *26*, 283–306.

Winkelmann, H. & Böhme, K. (2009). Anlage und Durchführung der Pilotierung der Bildungsstandards. In A. Bremerich-Vos, D. Granzer & O. Köller (Hrsg.), *Bildungsstandards Deutsch und Mathematik* (S. 32–43). Weinheim: Beltz.

Winkelmann, H. & Robitzsch, A. (2009). Modelle mathematischer Kompetenzen: Empirische Befunde zur Dimensionalität. In A. Bremerich-Vos, D. Granzer & O. Köller (Hrsg.), *Bildungsstandards Deutsch und Mathematik* (S. 169–196). Weinheim: Beltz.

Woods, C. M. (2006). Ramsay-curve item response theory (RC-IRT) to detect and correct for nonnormal latent variables. *Psychological Methods*, *11*(3), 253.

Wright, B. D. (1977). Misunderstanding the Rasch model – a heady tale. *Journal of Educational Measurement*, *14*, 219–225.

Wright, B. D. & Douglas, G. A. (1977). Best procedures for sample-free item analysis. *Applied Psychological Measurement*, *1*(2), 281–295.

Wu, H. & Browne, M. W. (2015). Quantifying adventitious error in a covariance structure as a random effect. *Psychometrika*, *80*, 571–600.

Wu, M. L. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement*, *29*, 15–27.

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. (2007). *ACER Conquest Version 2.0.* Mulgrave.

Wu, M. L., Douglas, A. & Monseur, C. (2002). *Issues in the design of the student assessment instruments for PISA 2000.* International Conference on Improving Surveys, Copenhagen.

Wuttke, J. (2007). Die Insignifikanz signifikanter Unterschiede: Der Genauigkeitsanspruch von PISA ist illusorisch. In T. Jahnke & W. Meyerhöfer (Hrsg.), *PISA & Co: Kritik eines Programms* (S. 99-246). Hildesheim: Franzbecker.

Xu, X. & von Davier, M. (2008). *Fitting the structured general diagnostic model to NAEP data* (ETS Research Report RR08-27). Princeton: ETS.

Xu, X. & von Davier, M. (2010). *Linking errors in trend estimation in large-scale surveys: a case study* (ETS Research Report RR10-10). Princeton: ETS.

Yen, W. M. (1984). Effects of local dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*, 125–145.

Yen, W. M. & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Hrsg.), *Educational measurement* (S. 111–154). Westport: Praeger Publisher.

Young, C. (2009). Model uncertainty in sociological research: An application to religion and economic growth. *American Sociological Review*, *74*, 380–397.

Yuan, K.-H. & Chan, W. (2008). Structural equation modeling with near singular covariance matrices. *Computational Statistics & Data Analysis*, *52*(10), 4842–4858.

Yucel, R. M. (2008). Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response. *Philosophical Transactions of the Royal Society A*, *366*, 2389–2404.

Zhang, J. & Stout, W. (1999). The theoretical DETECT index of multidimensionality and its application to approximate simple structure. *Psychometrika*, *64*, 213–249.

Zimmerman, D. L. & Nunez-Anton, V. A. (2009). *Antedependence models for longitudinal data.* CRC Press.

Zinbarg, R. E., Revelle, W., Yovel, I. & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω_h: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*, 123–133.

Zumbo, B. D. & Forer, B. (2011). Testing and measurement from a multilevel view: Psychometrics and validation. In J. A. Bouvaird, K. F. Geisinger & C. W. Buckendahl (Hrsg.), *High stakes testing in education – Science and practise in K-12 settings* (S. 177–190). New York: American Psychological Association.

Zwick, R. (1992). Statistical and psychometric issues in the measurement of educational achievement trends: Examples from the National Assessment of Educational Progress. *Journal of Educational Statistics*, *17*, 205–218.

Zwinderman, A. H. (1995). Pairwise parameter estimation in Rasch models. *Applied Psychological Measurement*, *19*, 369–375.